SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 1–18

GROWTH OF EDGE-HOMOGENEOUS TESSELLATIONS*

STEPHEN GRAVES†, TOMAŽ PISANSKI‡, AND MARK E. WATKINS†

Abstract. A *tessellation* is understood to be a 1-ended, locally finite, 3-connected planar map. The *edge-symbol* $\langle p, q; k, \ell \rangle$ of an edge of a tessellation T is a 4-tuple listing the valences p and q of its two incident vertices and the covalences k and ℓ of its two incident faces. To say that T is *edge-homogeneous* means that all edges of T have the same edge-symbol. By a result of Grünbaum and Shephard, each edge-transitive tessellation may be identified with its edge-symbol. It is shown that the growth rate of T is given by a function $g(t) = \frac{1}{2}(t - 2 + \sqrt{t^2 - 4t})$ of the single variable $t = (\frac{p+q}{2} - 2)(\frac{k+\ell}{2} - 2)$, except that the growth rate equals $g(t-1)$ when the edge-symbol of T or its planar dual has the form $\langle 3, q; 4, 4 \rangle$, where $q \geq 6$. Thus, for each integer $t \geq 4$, there are only finitely many edge-homogeneous tessellations whose growth rate equals $g(t)$, allowing a complete list of such tessellations to be compiled in terms of increasing growth rate. The maximum value of the quantity $\frac{1}{p} + \frac{1}{q} + \frac{1}{k} + \frac{1}{\ell}$ for tessellations with given value t is shown to decrease monotonically as t increases, while the minimum value decreases only asymptotically. Methods are demonstrated for concrete enumeration of the sets of faces and vertices at any given facial distance from a fixed face, edge, or vertex.

Key words. tessellation, edge-homogeneous, Bilinski diagram, exponential growth, generating function, transition matrix, eigenvalue

AMS subject classifications. 05B45, 05C12, 52C20, 15A18

DOI. 10.1137/070707026

1. Introduction.

In the present work, the term "tessellation" denotes an infinite, locally finite, 3-connected planar map that is *one-ended*, i.e., the deletion of no finite subgraph leaves more than a single infinite component. It is well known that any automorphism of the underlying graph of such a map is extendable to a homeomorphism of the plane [10]. If a tessellation is almost-transitive, then it is also dually locally finite; i.e., all facial walks are finite circuits (cf. [3, Theorem 2.3]). A tessellation is *edge-homogeneous* when there exists a 4-tuple $\langle p, q; k, \ell \rangle$ of integers ≥ 3, called the *edge-symbol* of the tessellation, such that for each edge, p and q are the valences of its two incident vertices, and k and ℓ are the covalences of its two incident faces. Grünbaum and Shephard [8] proved that edge-homogeneous tessellations are determined up to isomorphism by their edge-symbol and are, in fact, edge-transitive.

We determine the "growth rate" of edge-homogeneous tessellations outward from a central vertex, edge, or face, called its *root*. When F_n denotes the set of faces in the nth corona of a Bilinski diagram of a tesselation T, the growth rate is defined as

$$\gamma(T) = \lim_{n \to \infty} \sum_{j=0}^{n+1} |F_j| \Big/ \sum_{j=0}^{n} |F_j|,$$

if the limit exists; it is independent of the chosen root.

*Received by the editors October 29, 2007; accepted for publication (in revised form) May 3, 2008; published electronically October 24, 2008.
http://www.siam.org/journals/sidma/23-1/70702.html
†Department of Mathematics, Syracuse University, Syracuse, NY 13244-1150 (sgraves@syr.edu, mewatkin@syr.edu).
‡Department of Mathematics, IMFM, University of Ljubljana and University of Primorska, Jadranska 19, 1000 Ljubljana, Slovenia (Tomaz.Pisanski@fmf.uni-lj.si). Partially funded by ARRS grants M1-0160, M1-0176, M5-0164, L1-7230, and P1-0294.

For a rooted tessellation T with edge-symbol $\langle p, q; k, \ell \rangle$, the faces are of various "types" depending upon their covalence and their orientation with respect to the root. We obtain a system of recurrences which enable us to compute the numbers of faces of T of each type at regional distance $n+1$ (from the root) in terms of the number of each type at regional distance n ($n \in \mathbb{N}$). These recurrences yield a transition matrix A whose entries are multinomials in p, q, k, and ℓ. Using packages for symbolic computation, we can determine the spectrum of A explicitly for any edge-homogeneous tessellation. The ordinary generating function $\sum_{n\in\mathbb{N}} |F_n| z^n$ of the sequence $\{|F_n| : n \in \mathbb{N}\}$ is expressed in terms of A. Then we prove that the growth rate $\gamma(T)$ equals the eigenvalue of A of largest modulus (absolute value).

To an edge-homogeneous tessellation with edge-symbol $\langle p, q; k, \ell \rangle$ we associate a parameter t, which is a linear function of the product of the average valence $(p+q)/2$ and the average covalence $(k+\ell)/2$. The set of values assumed by t is exactly the set of integers ≥ 4. For each $t \geq 4$, let $\mathscr{T}(t)$ denote the set of edge-homogeneous tessellations associated with t. The sets $\mathscr{T}(t)$ are finite. Our main result is that $\gamma(T)$ for $T \in \mathscr{T}(t)$ is determined also by the single parameter t. Specifically, for each $t \geq 4$, if $T \in \mathscr{T}(t)$, then $\gamma(T) = g(t) = \frac{1}{2}\left(t - 2 + \sqrt{t^2 - 4t}\right)$ with the following exception: When the edge-symbol of T or its planar dual is of the form $\langle 3, q; 4, 4 \rangle$ for $q \geq 6$, then $\gamma(T) = g(t-1)$.

It is a folklore theorem that an edge-homogeneous tessellation T is finite, has quadratic growth, or has exponential growth (with respect to regional distance from a root) when the quantity $\mu(T) = \frac{1}{p} + \frac{1}{q} + \frac{1}{k} + \frac{1}{\ell}$ is > 1, $= 1$, or < 1, respectively. Denote by $m(t)$ and $M(t)$ the least and greatest values, respectively, of $\mu(T)$ for $T \in \mathscr{T}(t)$. We prove that $M(t)$ is strictly decreasing in t. The Lagrange multiplier method shows that $m(t)$ is asymptotic to $4/\left(2 + \sqrt{t}\right)$, although it is not monotonic.

Finally we demonstrate how to crunch some numbers to obtain exact values for the numbers of vertices, edges, and faces at any given facial distance from a central vertex, edge, or face.

This article considerably extends the work of Moran; in [11] she computed the growth rates of tessellations, in our notation, of the form $\langle p, p; k, k \rangle$ and determined when the limit $\lim_{n\to\infty} \sum_{j=0}^{n+1} |F_j| / \sum_{j=0}^{n} |F_j|$ exists for face-homogeneous triangulations of the hyperbolic plane.

2. Preliminaries. In order to give a precise definition of "growth rate," we use what may be called a *Bilinski diagram*. These diagrams were first used by Bilinski in his dissertation [1, 2] and more recently by Grünbaum and Shephard [9] and Moran [11].

DEFINITION 2.1. *Let M be a map that is rooted at some vertex x. Define a sequence of sets $\{U_n : n \geq 0\}$ of vertices and a sequence of sets $\{F_n : n \geq 0\}$ of faces as follows.*

- *Let $U_0 = \{x\}$, and let $F_0 = \emptyset$.*
- *For $n \geq 1$, let F_n denote the set of faces of M not in F_{n-1} that are incident with some vertex in U_{n-1}.*
- *For $n \geq 1$, let U_n denote the set of vertices of M not in U_{n-1} that are incident with some vertex in F_n.*

The stratification of M determined by $\{U_n\}$ and $\{F_n\}$ is called the Bilinski diagram *B of M* rooted at *v. In a similar way one can define a Bilinski diagram of M rooted at a face f. In this case $U_0 = \emptyset$ and $F_0 = \{f\}$. A Bilinski diagram is* concentric *if each subgraph $\langle U_n \rangle$ induced by U_n ($n \geq 1$) is a circuit. If a map yields a concentric*

Bilinski diagram regardless of which vertex or face is designated as its root, then the map is uniformly concentric.

Remark. In practice one may alter Definition 2.1 by letting the root be any vertex-induced, finite, simply connected submap.

Intuitively, one can label any planar map as a Bilinski diagram by arbitrarily selecting a vertex to comprise the singleton set U_0 and then calling by U_n each set of vertices on subsequent successive layers with (increasing) radius n. When the diagram is concentric, the layers induce "concentric" circuits $\langle U_n \rangle$. The annulus between two consecutive layers is partitioned by the set F_n of faces, which constitute the nth *corona*. Thus the vertices adjacent to a vertex in U_n lie in $U_{n-1} \cup U_n \cup U_{n+1}$, and the vertices incident with a face in F_n belong to $U_{n-1} \cup U_n$.

Let $\mathscr{M}_{a,b}$ denote the set of maps with all valences finite and $\geq a$ and all covalences finite $\geq b$. Let $\mathscr{M}_{a,b^+}$ denote the set of maps in $\mathscr{M}_{a,b}$ such that no two b-covalent faces are adjacent. Let $\mathscr{M}_{a^+,b}$ denote the set of maps in $\mathscr{M}_{a,b}$ such that no two a-valent vertices are adjacent.

The following proposition contains results from [12] and [4].

PROPOSITION 2.2. *Let the map M be labeled as a Bilinski diagram with respect to which $v \in U_m$ and $f \in F_n$, $(m, n \geq 1)$.*

(a) *If $M \in \mathscr{M}_{3,6} \cup \mathscr{M}_{3^+,5} \cup \mathscr{M}_{4,4} \cup \mathscr{M}_{5,3^+} \cup \mathscr{M}_{6,3}$, then M is uniformly concentric.*
(b) *If $M \in \mathscr{M}_{3,6} \cup \mathscr{M}_{3^+,5}$, then v is adjacent to at most one vertex in U_{m-1} and f is incident with at most two edges of $\langle U_{n-1} \rangle$.*
(c) *If $M \in \mathscr{M}_{4,4}$, then v is adjacent to at most one vertex in U_{m-1} and f is incident with at most one edge of $\langle U_{n-1} \rangle$.*
(d) *If $M \in \mathscr{M}_{5,3^+} \cup \mathscr{M}_{6,3}$, then v is adjacent to at most two vertices in U_{m-1} and f is incident with at most one edge of $\langle U_{n-1} \rangle$.*

The next proposition from [5] gives necessary conditions for uniform concentricity in terms of forbidden local configurations.

PROPOSITION 2.3. *If a map admits any of the following configurations, then it is not uniformly concentric:*

(a) *a* 3*-valent vertex incident with a* 3*-covalent face;*
(b) *a* 4*-valent vertex incident with two nonadjacent* 3*-covalent faces;*
(c) *an edge incident with two* 3*-valent vertices and two* 4*-covalent faces;*
(d) *a* 4*-covalent face incident with two nonadjacent* 3*-valent vertices;*
(e) *an edge incident with two* 4*-valent vertices and two* 3*-covalent faces.*

Note that these conditions are closed with respect to duality. Uniformly concentric tessellations form, in a sense, the general case. The nonconcentric Bilinski diagrams evidence some "closing up" at all but the first few levels, yielding a slower growth rate, but also requiring special computational considerations, as we will see in section 4.

DEFINITION 2.4. *Let the tessellation T be labeled as a Bilinski diagram. The* growth rate *of T is defined as*

$$\gamma(T) = \lim_{n \to \infty} \sum_{j=0}^{n+1} |F_j| \Big/ \sum_{j=0}^{n} |F_j|$$

when this limit exists and is finite.

It is not hard to show that the growth rate γ just defined is equal to the growth rate defined in terms of the standard distance metric $d(-,-)$, provided that the covalences of the map are not arbitrarily large. Consider a Bilinski diagram of a map M with maximum covalence k, and let the root be a vertex x. Let y be an arbitrary

vertex in U_{n+1} of the Bilinski diagram. Since $1 \leq d(y, U_n) \leq \lfloor k/2 \rfloor$, one easily obtains by induction that

$$n \leq d(x, y) \leq n \lfloor k/2 \rfloor .$$

Hence every n-ball with respect to $d(-,-)$ centered at x is contained in the union of the first n layers of vertices of the Bilinski diagram, while U_n is contained in the $(n\lfloor k/2 \rfloor)$-ball centered at x of the underlying graph.

To place the notion of homogeneity in a more general context, let us make the following definitions (cf. [14]):

DEFINITION 2.5. *Let T be a tessellation, and let f be a k-covalent face of T. A* valence sequence at f *is a cyclic k-tuple $(p_1, p_2, \ldots, p_k)$ of integers ≥ 3 that lists in cyclic order the valences of the vertices incident with f as one proceeds around f in either the clockwise or counterclockwise direction. If a given cyclic k-tuple is a valence sequence of every face of T, then we say that T is* face-homogeneous.

If in this definition we swap the words "vertex" and "face" and the words "valence" and "covalence," then we will have defined a *vertex-homogeneous* map.

Suppose that T is edge-homogeneous with edge-symbol $\langle p, q; k, \ell \rangle$. If $p = q$, then T is vertex-homogeneous with covalence sequence $(k, \ell, k, \ell, \ldots, k, \ell)$. Dually, if $k = \ell$, then T is face-homogeneous with valence sequence $(p, q, p, q, \ldots, p, q)$. Clearly a map that is both p-valent and k-covalent is vertex-, face-, and edge-homogeneous. While each permissible edge-symbol determines a unique edge-transitive map (by Proposition 2.8 below), a covalence sequence may be realized by infinitely or finitely many vertex-homogeneous maps or by no map at all, and any map so determined may or may not be vertex-transitive. (This question is the subject of [14].)

PROPOSITION 2.6 (Moran [11], Theorems 7.1 and 9.1). *Let T be a vertex-homogeneous tessellation whose planar dual is T^*.*

(a) *If $\gamma(T)$ exists, then so does $\gamma(T^*)$, and $\gamma(T^*) = \gamma(T)$.*
(b) *The recurrences that determine $\gamma(T)$ are independent of the root of the Bilinski diagram used to compute them.*

The following result will be used for a special case in section 4.

PROPOSITION 2.7 (Moran [11], pp. 159, 163). *Let T be a p-valent, k-covalent tessellation, where $1/p + 1/k \leq 1/2$.*

(a) *If $k \geq 4$, then its growth rate is given by*

$$\gamma(T) = \frac{(kp - 2p - 2k + 2) + \sqrt{(kp - 2p - 2k + 2)^2 - 4}}{2}. \tag{2.1}$$

(b) *If $k = 3$ and $p \geq 7$, then its growth rate is given by*

$$\gamma(T) = \frac{p - 4 + \sqrt{(p-4)^2 - 4}}{2}. \tag{2.2}$$

(c) *If $1/p + 1/k < 1/2$, then $\gamma(T)$ is an irrational number > 1.*

Note that if the parameters for any of the three regular Euclidean tessellations (where $1/p + 1/k = 1/2$) are substituted into (2.1), then we obtain $\gamma = 1$. In this same work [11, Theorem 6.1], Moran also determined the growth rates of all 3-covalent face-homogeneous maps and found $\lim_{n\to\infty} \sum_{j=0}^{n+1} |F_j| / \sum_{j=0}^{j=n} |F_j|$ to exist in all cases except when the valence sequence has the form $(2j_1, 2j_2, 4)$, where $j_1 \neq j_2$.

For an edge-homogeneous tessellation T with edge-symbol $\langle p, q; k, \ell \rangle$, we define

$$\mu(T) = \frac{1}{p} + \frac{1}{q} + \frac{1}{k} + \frac{1}{\ell}.$$

It is well known that if $\mu(T) > 1$, then T is finite—if it is realizable at all—and if so, then it has a "normal" realization on the sphere, in the sense that faces of equal covalence are congruent regular polygons. If $\mu(T) = 1$ and if T is realizable, then T has a normal realization in the Euclidean plane, $\sum_{j=0}^{n} |F_j|$ grows quadratically in n, and $\gamma(T) = 1$. However, if $\mu(T) < 1$, then T has a normal realization in the hyperbolic plane, $\sum_{j=0}^{n} |F_j|$ grows exponentially in n, and $\gamma(T) > 1$.

For the tessellations considered in this article, the following strong result was obtained in [8].

PROPOSITION 2.8 (Grünbaum and Shephard [8]). *Let p, q, k, ℓ be integers ≥ 3. There exists an edge-homogeneous, 3-connected, finite or 1-ended map with edge-symbol $\langle p, q; k, \ell \rangle$ if and only if exactly one of the following holds:*

(a) *all of $p, q, k,$ and ℓ are even;*
(b) *$k = \ell$ is even and at least one of p, q is odd;*
(c) *$p = q$ is even and at least one of k, ℓ is odd;*
(d) *$p = q$, $k = \ell$, and all are odd.*

Such a map is edge-transitive and is uniquely determined (up to isomorphism) by its edge-symbol. If $p = q$, then it is vertex-transitive. If $k = \ell$, then it is face-transitive. Finally, the parameters p, q, k, ℓ determine the map up to homeomorphism of the plane.

We remark that, for some edge-symbols, there exist more than one *multi*-ended map with that edge-symbol. A detailed classification of all edge-transitive planar maps is found in [7].

3. The generating function. Let a tessellation T with edge-symbol $\langle p, q; k, \ell \rangle$ be labeled in accordance with a Bilinski diagram. If the root is a face, then $|F_0| = 1$. Otherwise $F_0 = \emptyset$. If the root is a vertex x, then $U_0 = \{x\}$ and $|F_1|$ equals the valence of x. If the root is an edge, we let U_0 consist of its two incident vertices, while F_1 consists of all faces incident with one or both of these two vertices.

Suppose that the set F_n is partitioned into m *types* of faces; the type of a face $f \in F_n$ is determined by its covalence, the number of vertices incident with f of each valence that lie in U_{n-1}, and the number of vertices incident with f of each valence that lie in U_n. Let the column vector $\mathbf{v}_n = [v_1, \ldots, v_m]^t$ list the number of faces in F_n of each type. Suppose further that for all $n \geq 1$ and each $i, j \in \{1, \ldots, m\}$, there exists a constant $a_{i,j}$ denoting the number of faces in F_{n+1} of the ith type whose existence is due to each face in F_n of the jth type. The existence of such constants $a_{i,j}$ for edge-homogeneous tessellations will be demonstrated by direct computation in section 4.

The $m \times m$ matrix $A = [a_{i,j}]$, called the *transition matrix*, satisfies the recurrence

$$\mathbf{v}_{n+1} = A\mathbf{v}_n, \quad n \geq 1.$$

Each entry of A and of $\mathbf{v}_n$ is a multinomial in some or all of $p, q, k,$ and ℓ. Let the row vector $\mathbf{j} = [1, 1, \ldots, 1]$ be regarded as an $(m \times 1)$-matrix. Then $|F_{n+1}| = \mathbf{j}A\mathbf{v}_n = \mathbf{j}A^n\mathbf{v}_0$, by induction for any $n \geq 0$ once the initial condition $\mathbf{v}_0$ is given. In practice, however, the vector $\mathbf{v}_0$ is fictitious, because the types of faces that need to be counted in $\mathbf{v}_0$ generally never occur in the nth corona when $n \geq 1$. For example, in the proof of Lemma 4.3 below, if the root is a 3-valent vertex together with its three incident faces, then F_0 consists of three faces of a type that cannot exist elsewhere in the Bilinski diagram and hence does not appear in the list of face-types. Our mechanism for dealing with such situations is to replace $A^n\mathbf{v}_0$ by $A^{n-1}\mathbf{v}_1$ for $n \geq 1$.

We thus compute the ordinary generating function $\varphi(z)$ of the sequence $\{|F_n| : n \geq 0\}$:

$$\begin{aligned}\varphi(z) &= \sum_{n=0}^{\infty} |F_n| z^n \\ &= |F_0| + \sum_{n=1}^{\infty} \left(\mathbf{j} A^{n-1} \mathbf{v}_1\right) z^n \\ &= |F_0| + z\mathbf{j} \left[\sum_{n=0}^{\infty} (zA)^n\right] \mathbf{v}_1\end{aligned}$$

from which the next theorem follows.

THEOREM 3.1. *The ordinary generating function for the number $|F_n|$ of faces in the nth corona of a concentric Bilinski diagram of an edge-homogeneous tessellation with transition matrix A is*

$$\varphi(z) = |F_0| + z\mathbf{j}(I - zA)^{-1}\mathbf{v}_1,$$

where A is the transition matrix, $\mathbf{j}$ is the row vector of 1s, and $\mathbf{v}_1$ is a column vector listing the distribution of face-types in the first corona F_1.

The following result is from [13, p. 159].

PROPOSITION 3.2. *Assume that a rational generating function $u(z)/v(z) = \sum a_n z^n$, with $u(z)$ and $v(z)$ relatively prime and $v(0) \neq 0$, has a unique pole $1/\beta$ of smallest modulus, and let its multiplicity be m. Then*

$$a_n = C\beta^n n^{m-1} + o(\beta^n n^{m-1}), \quad \text{where} \quad C = m\frac{(-\beta)^m f(1/\beta)}{v^{(m)}(1/\beta)}.$$

LEMMA 3.3. *Let $u(z)/v(z)$ be a rational generating function for $\sum a_n z^n$ such that $v(z)$ has a unique root of smallest modulus $1/\lambda$ and $v(0) \neq 0$. Then*

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lambda.$$

Proof. Letting m be the multiplicity of the root of $v(z)$ at $1/\lambda$, we have from Sedgewick and Flajolet's proof of Proposition 3.2 that

$$a_n \sim \frac{c_0}{(m-1)!} n^{m-1}\lambda^n$$

for some nonzero constant c_0. This immediately gives

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lambda. \qquad \square$$

For a matrix B, let $\mathrm{cof}(B)$ denote the matrix whose (i,j)-entry is the cofactor of the (i,j)-entry of B, and let $\chi(B)$ denote its characteristic polynomial.

THEOREM 3.4. *If the $m \times m$ transition matrix A of a tessellation T has a unique eigenvalue $\lambda > 1$ of largest modulus, then the growth rate of T is λ.*

Proof. By Theorem 3.1,

$$\varphi(z) = \sum_{n=0}^{\infty} |F_n| z^n = |F_0| + z\left[\mathbf{j}(I - zA)^{-1}\mathbf{v_1}\right].$$

If $\varphi(z)$ is written as a rational function $u(z)/v(z)$, then $v(z)$ is determined by $(I - zA)^{-1}$. Specifically,

$$\begin{aligned}(I - zA)^{-1} &= \frac{1}{\det(I - zA)}\operatorname{cof}(I - zA)\\ &= \frac{1}{(-z)^m \det(A - \frac{1}{z}I)}\operatorname{cof}(I - zA)\\ &= \frac{1}{(-z)^m \chi(\frac{1}{z})}\operatorname{cof}(I - zA).\end{aligned}$$

Since elements of $\operatorname{cof}(I - zA)$ are polynomials in z, the denominator of $\varphi(z)$ is of the form $v(z) = (-z)^m f(1/z)$; $\chi(1/z)$ is a polynomial of degree m in $1/z$. Hence $v(z)$ has nonzero constant term. This in turn gives that the roots of $v(z)$ occur precisely at the reciprocals of nonzero eigenvalues of A, and so the root of minimum modulus of $v(z)$ is $1/\lambda$.

Let $\psi(z)$ be the generating function of the sequence $\{\sum_{j=0}^{n} |F_j| : n \geq 0\}$. Then $\psi(z) = \sum_{n=0}^{\infty}(\sum_{j=0}^{n} |F_m|)z^n = \sum_{n=0}^{\infty} |F_n|z^n/(1-z) = \varphi(z)/(1-z)$. The denominator of $\psi(z)$ is $(1 - z)v(z)$, which has no additional root of modulus less than $1/\lambda$. Hence by Lemma 3.3,

$$\gamma(T) = \lim_{n\to\infty} \sum_{j=0}^{n+1} |F_j| \Big/ \sum_{j=0}^{n} |F_j|. \qquad \square$$

4. The growth formula. Our first main result is the following.

THEOREM 4.1. *Let the function $g : \{t \in \mathbb{Z} : t \geq 4\} \to [1, \infty)$ be given by*

$$g(t) = \frac{1}{2}\left(t - 2 + \sqrt{t^2 - 4t}\right). \tag{4.1}$$

Let T be an edge-homogeneous tessellation with edge-symbol $\langle p, q; k, \ell\rangle$, and let

$$t = \left(\frac{p+q}{2} - 2\right)\left(\frac{k+\ell}{2} - 2\right). \tag{4.2}$$

Then exactly one of the following holds:

(a) *the growth rate of T is $g(t)$; or*
(b) *the edge-symbol of T or its planar dual is $\langle 3, q; 4, 4\rangle$ for some $q \geq 6$, and the growth rate of T is $g(t - 1)$.*

The proof of the theorem is embodied in four lemmas which partition the possibilities for the edge-symbol. Case (a) is realized by each of the first three of these lemmas, and all of the associated Bilinski diagrams are uniformly concentric. The fourth lemma realizes Case (b), where the associated Bilinski diagram is not concentric. Some of the eigenvalues in the proofs of these lemmas were obtained using *Maple*.

To fix notation for all four lemmas, we assume that T is an edge-homogeneous tessellation with edge-symbol $\langle p, q; k, \ell\rangle$, that g is given by (4.1), and that t is given by (4.2). The average valence is $r = (p+q)/2$, and the average covalence is $s = (k+\ell)/2$. Hence by (4.2), $t = (r - 2)(s - 2)$.

Remark. Observe that

- $g(t) = \frac{t}{2}(1 - \frac{2}{t} + \sqrt{1 - \frac{4}{t}})$, and so $\lim_{t\to\infty} \frac{g(t)}{t} = 1$;
- $\frac{1}{g(t)} = \frac{1}{2}(t - 2 - \sqrt{t^2 - 4t})$.

The most general case is treated in the first lemma.

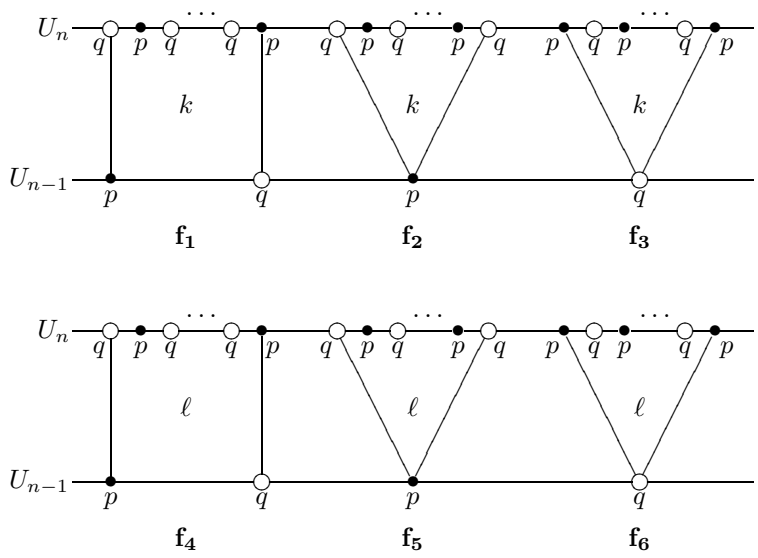


FIG. 1. *Face types for Lemma* 4.2*, Case* 1.

LEMMA 4.2. *If* p, q, k*, and* ℓ *are all at least* 4*, then* $\gamma(T) = g(t)$.

Proof. By assumption, $T \in \mathscr{M}_{4,4}$. Hence by Proposition 2.2(a), T is uniformly concentric. The proof of this lemma is broken into four cases corresponding to the four cases listed in Proposition 2.8.

Case 1: *all of* p, q, k*, and* ℓ *are even.*

Remark. This is the most complicated case. By explaining our procedure in considerable detail in this case, we hope to omit much of the detail in the subsequent, simpler cases and in the other lemmas of this section.

We assume the sets of vertices and faces of T to be labeled with respect to a concentric Bilinski diagram. By Proposition 2.2(c), there can be up to six *types* of faces. For each $n \geq 2$, the faces in the nth corona F_n have the following descriptions, respectively, and are illustrated in Figure 1:

Type $\mathbf{f_1}$ is a k-covalent face incident with one edge in U_{n-1};

Type $\mathbf{f_2}$ is a k-covalent face incident with exactly one p-valent vertex in U_{n-1};

Type $\mathbf{f_3}$ is a k-covalent face incident with exactly one q-valent vertex in U_{n-1};

Type $\mathbf{f_4}$ is an ℓ-covalent face incident with one edge in U_{n-1};

Type $\mathbf{f_5}$ is an ℓ-covalent face incident with exactly one p-valent vertex in U_{n-1};

Type $\mathbf{f_6}$ is an ℓ-covalent face incident with exactly one q-valent vertex in U_{n-1}.

Suppose that $\mathbf{v_n}$ denotes the column vector that lists the number of faces in F_n of each of these six types. The transition matrix $A = [a_{i,j}]$ is then a (6×6)-matrix that satisfies

$$\mathbf{v}_{n+1} = A\mathbf{v}_n, \quad n \geq 1. \tag{4.3}$$

The way that the entries of $a_{i,j}$ are obtained is indicated by Figure 2. We understand that if f_1 and f_2 are adjacent faces in F_n and $g \in F_{n+1}$ is adjacent to neither f_1 nor f_2 but shares an incident vertex with both of them, then each of f_1 and f_2 is given half-credit for the existence of g.

Let us, by way of an example, compute $a_{2,1}$, the number of Type $\mathbf{f_2}$ faces in F_{n+1} produced by each Type $\mathbf{f_1}$ face $f \in F_n$. The face f is incident with $\frac{1}{2}(k-2)$ p-valent

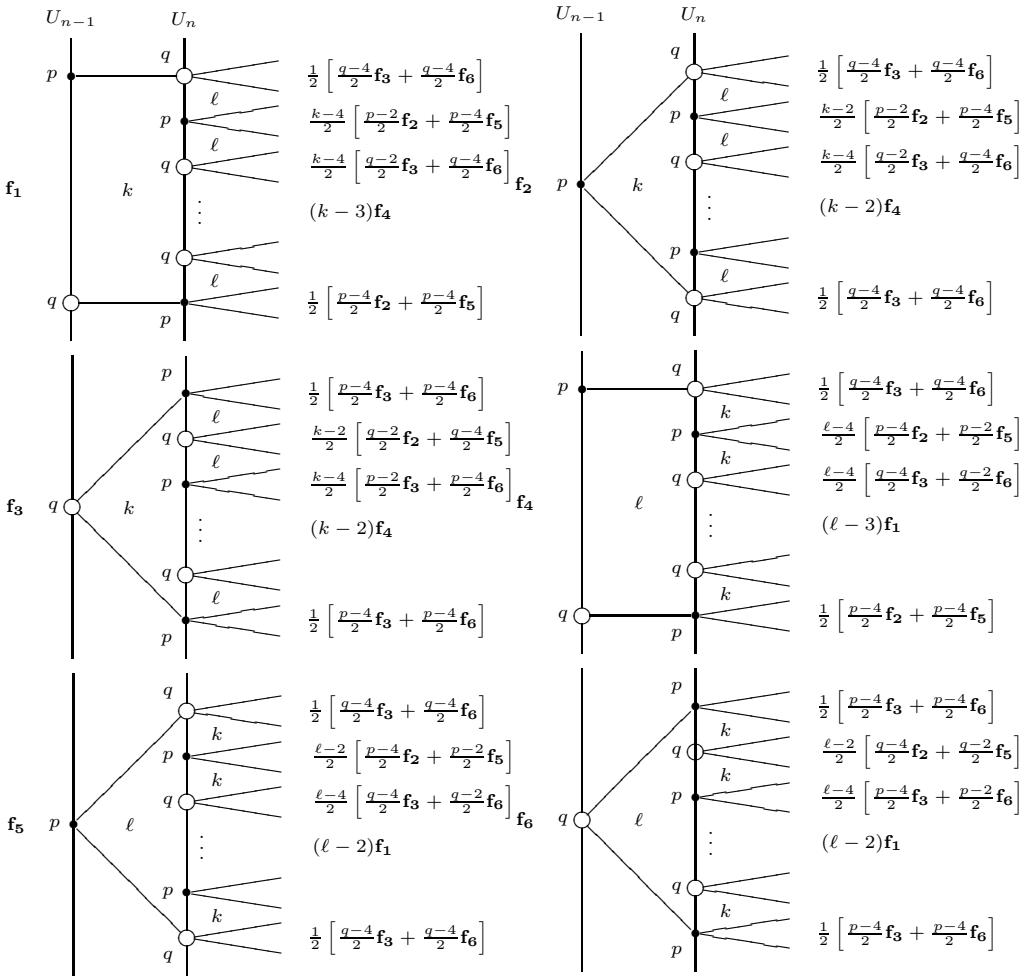


FIG. 2. *"Offspring" of various face types for Lemma* 4.2*, Case* 1.

vertices in U_n. Each of exactly $\frac{1}{2}(k-4)$ of these p-valent vertices is adjacent to $p-3$ vertices in U_{n+2} and is therefore incident with $\frac{1}{2}(p-2)$ Type $\mathbf{f_2}$ faces in F_{n+1}. (Recall that every edge is incident with one k-covalent face and one ℓ-covalent face.) The one remaining p-valent vertex is incident with $\frac{1}{2}(p-4)$ Type $\mathbf{f_2}$ faces in F_{n+1}. Since this vertex is also incident with another face in F_n, we count only half of its contribution. As a total, we get

$$a_{2,1} = \frac{k-4}{2}\cdot\frac{p-2}{2} + \frac{1}{2}\cdot\frac{p-4}{2} = \frac{1}{4}(pk-3p-2k+4).$$

In this manner, one obtains all 36 entries of the following transition matrix:

$$A = \begin{pmatrix} 0 & 0 & 0 & \ell-3 & \ell-2 & \ell-2 \\ \frac{kp-3p-2k+4}{4} & \frac{(k-2)(p-2)}{4} & \frac{kp-2p-2k}{4} & \frac{(\ell-3)(p-4)}{4} & \frac{(\ell-2)(p-4)}{4} & \frac{\ell p-2p-4\ell+8}{4} \\ \frac{kq-3q-2k+4}{4} & \frac{kq-2q-2k}{4} & \frac{(k-2)(q-2)}{4} & \frac{(\ell-3)(q-4)}{4} & \frac{\ell q-2q-4\ell+8}{4} & \frac{(\ell-2)(q-4)}{4} \\ k-3 & k-2 & k-2 & 0 & 0 & 0 \\ \frac{(k-3)(p-4)}{4} & \frac{(k-2)(p-4)}{4} & \frac{kp-2p-4k+8}{4} & \frac{\ell p-3p-2\ell+4}{4} & \frac{(\ell-2)(p-2)}{4} & \frac{\ell p-2p-2\ell}{4} \\ \frac{(k-3)(q-4)}{4} & \frac{kq-2q-4k+8}{4} & \frac{(k-2)(q-4)}{4} & \frac{\ell q-3q-2\ell+4}{4} & \frac{\ell q-2q-2\ell}{4} & \frac{(\ell-2)(q-2)}{4} \end{pmatrix}.$$

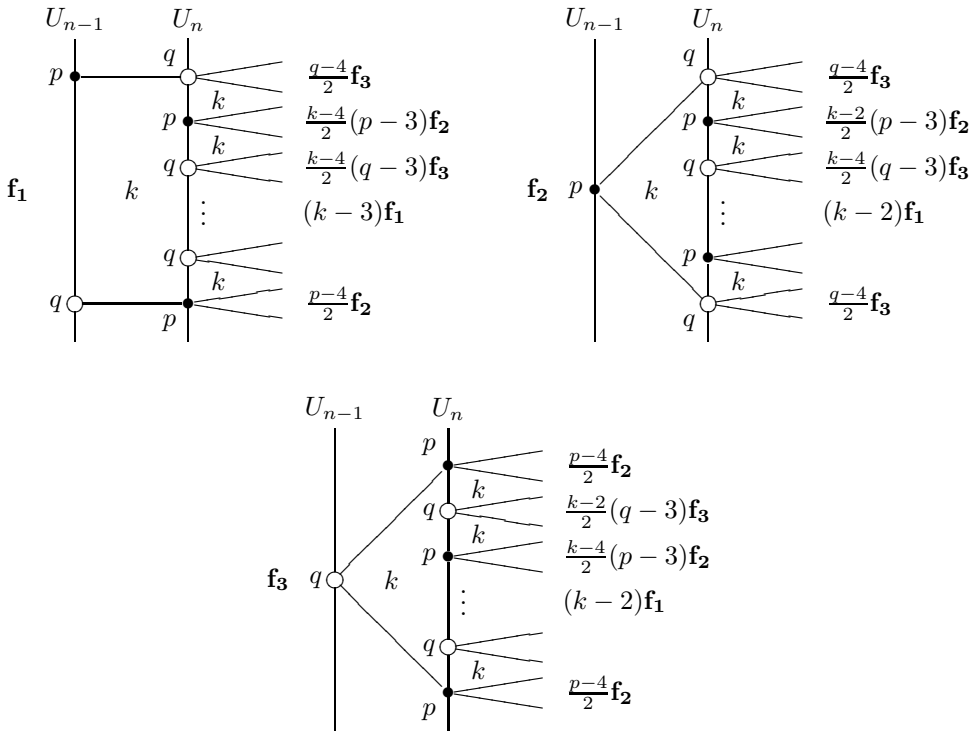


FIG. 3. *"Offspring" for face types for Lemma* 4.2*, Case* 2.

The eigenvalue of A with maximum modulus is

$$\begin{aligned}
&\frac{1}{8}\left(pk+p\ell+qk+q\ell-4p-4q-4k-4\ell+8\right)\\
&\qquad+\frac{1}{8}\sqrt{(p+q-4)\left(k+\ell-4\right)\left(pk+p\ell+qk+q\ell-4p-4q-4k-4\ell\right)}\\
&=\frac{1}{8}\left[(p+q)\left(k+\ell\right)-4(p+q+k+l)+8\right]\\
&\qquad+\frac{1}{8}\sqrt{(p+q-4)\left(k+\ell-4\right)\left[(p+q)\left(k+\ell\right)-4(p+q+k+\ell)\right]}\\
&=\frac{1}{2}\left[rs-2(r+s)+2+\sqrt{(r-2)(s-2)\left[rs-2(r+s)\right]}\right]\\
&=\frac{1}{2}\left[(r-2)(s-2)-2+\sqrt{(r-2)(s-2)\left[(r-2)(s-2)-4\right]}\right]\\
&=\frac{1}{2}\left(t-2+\sqrt{t^2-4t}\right)=g(t).
\end{aligned}$$

By Theorem 3.4, $\gamma(T)=g(t)$ as claimed.

Case 2: $k=\ell$ *and at least one of* p,q *is odd.*

This case may be considered as a special case of Case 1, where T has edge-symbol $\langle p,q;k,k\rangle$. We do not actually use that p or q is odd, but for the matrix entries to make sense, k must be even. Here the face Types $\mathbf{f_1}$ and $\mathbf{f_4}$ of the previous case are identified, as are Types $\mathbf{f_2}$ and $\mathbf{f_5}$, as well as Types $\mathbf{f_3}$ and $\mathbf{f_6}$, and their "offspring" are as seen in Figure 3.

The transition matrix is thus the following (3×3)-matrix:

$$A = \begin{bmatrix} k-3 & k-2 & k-2 \\ \frac{1}{2}(kp-3k-3p+8) & \frac{1}{2}(k-2)(p-3) & \frac{1}{2}(kp-3k-2p+4) \\ \frac{1}{2}(kq-3k-3q+8) & \frac{1}{2}(kq-3k-2q+4) & \frac{1}{2}(k-2)(q-3) \end{bmatrix}.$$

Again letting $r = \frac{p+q}{2}$, $s = \frac{k+\ell}{2}$, and $t = (r-2)(s-2)$, we compute the eigenvalue of A with maximum modulus.

$$\begin{aligned} &\frac{1}{4}\left[(p+q)k - 2(p+q) - 4k + 4\sqrt{(p+q-4)(k-2)((p+q)k-2(p+q)-4k)}\right] \\ &= \frac{1}{2}\left[(r-2)(s-2) - 2 + \sqrt{(r-2)(s-2)\left[(r-2)(s-2)-4\right]}\right] \\ &= \frac{1}{2}\left(t - 2 + \sqrt{t^2-4t}\right) = g(t). \end{aligned}$$

Case 3: *$p = q$ is even, and at least one of k, ℓ is odd.*

In this case the tessellation is the planar dual of a tessellation described by Case 2. By Proposition 2.6(a), the growth rate is the same as that of its dual.

Case 4: *$p = q$, $k = \ell$, and all are odd.*

By Proposition 2.7(a), the growth rate of T is given by (2.1):

$$\begin{aligned} \gamma(T) &= \frac{1}{2}\left[pk - 2p - 2k + 2 + \sqrt{(pk-2p-2k-2)^2 - 4}\right] \\ &= \frac{1}{2}\left[(p-2)(k-2) - 2 + \sqrt{\left[(p-2)(k-2)-2\right]^2 - 4}\right]. \end{aligned}$$

Trivially, p is the average valence, and k is the average covalence. With $t = (p-2)(k-2)$, we obtain

$$\gamma(T) = \frac{1}{2}\left(t - 2 + \sqrt{t^2-4t}\right) = g(t). \qquad \square$$

The remaining three lemmas exhaust the special cases when T is not in $\mathscr{M}_{4,4}$. For the first of these, the edge-symbol of T is $\langle 3, q; k, k\rangle$. The average valence is $r = \frac{1}{2}(3+q)$ and the average covalence is $s = k$, and so $t = \frac{1}{2}(q-1)(k-2)$.

LEMMA 4.3. *If T has edge-symbol $\langle 3, q; k, k\rangle$, where $q \geq 4$ and $k \geq 6$, then $\gamma(T) = g(t)$.*

Proof. By Proposition 2.8, k must be even. Since $T \in \mathscr{M}_{3,6}$, T is uniformly concentric by Proposition 2.2(a), and we assume that the sets of vertices and faces of T have been labeled consistently with a Bilinski diagram. By Proposition 2.2(b), except perhaps in F_0 or F_1, T admits only three types of faces; for $n \geq 2$, a face $f \in F_n$ is of the following type:

Type $\mathbf{f_1}$ if f is incident with exactly one edge in the subgraph induced by U_{n-1};

Type $\mathbf{f_2}$ if f is incident with exactly two adjacent edges in the subgraph induced by U_{n-1};

Type $\mathbf{f_3}$ if f is incident with exactly one q-valent vertex in U_{n-1}.

In the instance of a Type $\mathbf{f_2}$ face, the "two adjacent edges" are incident with a common 3-valent vertex. Figure 4 shows the "offspring" of these three face types. The transition matrix A for this lemma is the following (3×3)-matrix:

$$A = \begin{bmatrix} k-4 & k-6 & k-4 \\ 1/2 & 1 & 1 \\ \frac{1}{2}(kq-3k-3q+8) & \frac{1}{2}(k-4)(q-3) & \frac{1}{2}(k-2)(q-3) \end{bmatrix}.$$

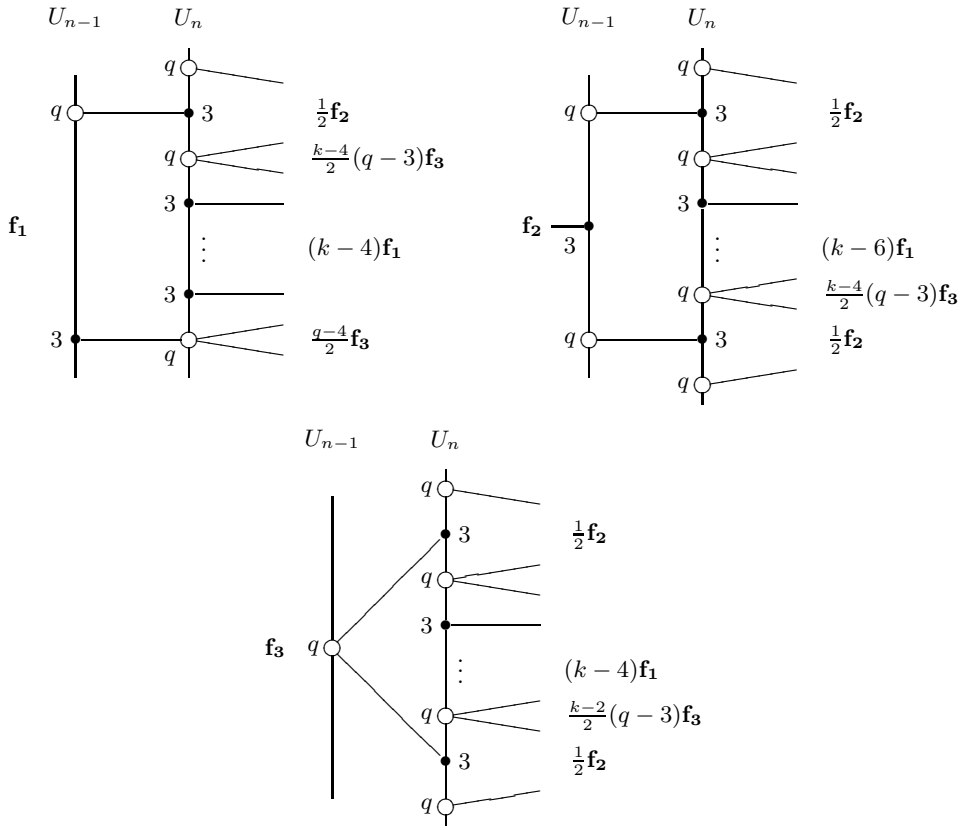


FIG. 4. *"Offspring" for face types for Lemma* 4.3.

The eigenvalue of A of maximum modulus is

$$\begin{aligned}
&\frac{1}{4}\left[qk-2q-k-2+\sqrt{(q-1)(k-2)(qk-2q-k-6)}\right]\\
&=\frac{1}{2}\left[\frac{1}{2}(q-1)(k-2)-2+\sqrt{\frac{1}{2}(q-1)(k-2)\left[\frac{1}{2}(q-1)(k-2)-4\right]}\right]\\
&=\frac{1}{2}\left(t-2+\sqrt{t^2-4t}\right)=g(t). \qquad \square
\end{aligned}$$

LEMMA 4.4. *If T or its planar dual has edge-symbol $\langle 3,3;k,k\rangle$, where $k\geq 6$, then $\gamma(T)=g(t)$.*

Proof. Clearly $t=k-2$. By Proposition 2.7(b), we use (2.2) to obtain

$$\begin{aligned}
\gamma(T)&=\frac{1}{2}\left(k-4+\sqrt{(k-4)^2-4}\right)\\
&=\frac{1}{2}\left(t-2+\sqrt{t^2-4t}\right)=g(t). \qquad \square
\end{aligned}$$

The one remaining class of edge-symbols to be considered is that of the form $\langle 3,q;4,4\rangle$, where $q\geq 6$. The average valence is $r=\frac{1}{2}(3+q)$, and the average covalence is trivially $s=4$. Thus $t=(r-2)(s-2)=q-1$. As previously remarked, this is the one situation where no Bilinski diagram of T or of its planar dual is concentric (cf.

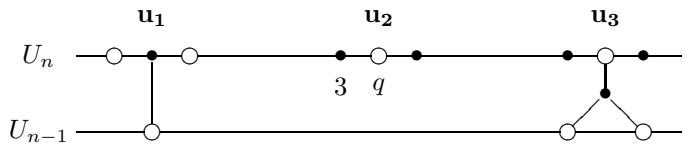


FIG. 5. *Vertex types for Lemma* 4.5.

Proposition 2.3(d)). The concurrence of small values in the edge-symbol causes some "closing up" and hence a slightly slower growth rate than for the other tessellations with the same t-value. It is notable that the class of edge-homogeneous planar maps with edge-symbol of the form $\langle 3, q; 4, 4\rangle$ and their planar duals is exactly the class of edge-transitive planar maps with the property that no Petrie walk is a double ray. Rather, all Petrie walks are circuits of length $2q$ (see [7], Theorem 6.3).

LEMMA 4.5. *If T has edge-symbol $\langle 3, q; 4, 4\rangle$, where $q \geq 6$, then $\gamma(T) = g(t-1)$.*

Proof. By Proposition 2.6(a), recurrences using vertices instead of faces yield the same growth rate. Except perhaps in the U_0 or U_1, T admits only the following three types of vertices (see Figure 5). For $n \geq 2$, a vertex $u \in U_n$ is of the following type:

Type $\mathbf{u_1}$ if u is 3-valent and has exactly one neighbor in U_{n-1};

Type $\mathbf{u_2}$ if u is q-valent, has no neighbor in U_{n-1} and two neighbors in U_n;

Type $\mathbf{u_3}$ if u is one of a pair of adjacent vertices of which the q-valent vertex has no neighbor in U_{n-1} and three neighbors in U_n, while the 3-valent member of the pair has two neighbors in U_{n-1}.

The two Type $\mathbf{u_3}$ vertices are treated as a single item in the recurrence computation, and therefore one must remember to double the number of Type $\mathbf{u_3}$ vertices when enumerating the sets U_n. The transition matrix A for this lemma is the (3×3)-matrix:

$$A = \begin{bmatrix} 0 & q-4 & q-5 \\ 0 & q-5 & q-6 \\ 0 & 1 & 1 \end{bmatrix}.$$

The eigenvalue of A with the largest modulus is

$$\frac{1}{2}\Big(q-4+\sqrt{(q-4)^2-4}\Big) = \frac{1}{2}\Big(t-3+\sqrt{(t-3)^2-4}\Big) = g(t-1). \qquad \square$$

Note that when $q = 6$, we have the familiar Euclidean tessellation with rhombi (designated by Coxeter [6, p. 61] as $[2\{6,3\}]\,\{6,3\}$), and the characteristic polynomial of A factors as $(x+1)^2(x-1)^2$. Setting $q = 7$ gives the "first" hyperbolic map of this kind; the roots are -1 and $\frac{1}{2}(3 \pm \sqrt{5})$.

5. Some comparisons. A lot of numerical data can be generated from the formulas of the preceding section. It is useful to have an idea of their orders of magnitude. In Table 1, we list the tessellations T, identified by their edge-symbol, having a t-value of $t \leq 9$, sorted first by increasing growth $\gamma(T)$, then by increasing t-value, and thirdly by decreasing value of $\mu(T)$. We also include the number of the lemma in the preceding section that gives the first two values. In the case of Lemma 4.2, the integer in parentheses indicates the appropriate case within the proof. To reduce redundant information, for tessellations that are not self-dual we have not listed both the tessellation and its dual. In particular, for a pair of dual tessellations covered by Cases 2 and 3 of Lemma 4.2, we list only one of the two tessellations.

Remark. We see from Table 1 that the edge-homogeneous tessellations with the slowest exponential growth rate have growth rate $\frac{1}{2}(3+\sqrt{5}) \approx 2.618$. There exist

TABLE 1
Edge-homogeneous tessellations with $4 \leq t \leq 9$.

$\gamma(T)$	t	$\mu(T)$	Edge-symbol	Lemma
1	4	1	$\langle 3,3;6,6\rangle$	3.4
1	4	1	$\langle 4,4;4,4\rangle$	3.2(1)
1	5	1	$\langle 3,6;4,4\rangle$	3.5
2.618	5	0.9524	$\langle 3,3;7,7\rangle$	3.4
2.618	5	0.95	$\langle 4,5;4,4\rangle$	3.2(2)
2.618	6	0.9762	$\langle 3,7;4,4\rangle$	3.5
3.7321	6	0.9167	$\langle 3,3;8,8\rangle$	3.4
3.7321	6	0.9167	$\langle 3,4;6,6\rangle$	3.3
3.7321	6	0.9167	$\langle 4,6;4,4\rangle$	3.2(1)
3.7321	6	0.9	$\langle 4,5;4,4\rangle$	3.2(2)
3.7321	7	0.9583	$\langle 3,8;4,4\rangle$	3.5
4.7913	7	0.8929	$\langle 4,7;4,4\rangle$	3.2(2)
4.7913	7	0.8889	$\langle 3,3;9,9\rangle$	3.4
4.7913	7	0.8667	$\langle 4,4;5,6\rangle$	3.2(3)
4.7913	8	0.9444	$\langle 3,9;4,4\rangle$	3.5
5.8284	8	0.875	$\langle 4,8;4,4\rangle$	3.2(1)
5.8284	8	0.8667	$\langle 3,5;6,6\rangle$	3.3
5.8284	8	0.8667	$\langle 3,3;10,10\rangle$	3.4
5.8284	8	0.8429	$\langle 5,7;4,4\rangle$	3.2(2)
5.8284	8	0.8333	$\langle 4,4;6,6\rangle$	3.2(1)
5.8284	9	0.9333	$\langle 3,10;4,4\rangle$	3.5
6.8541	9	0.8611	$\langle 4,9;4,4\rangle$	3.2(2)
6.8541	9	0.8485	$\langle 3,3;11,11\rangle$	3.4
6.8541	9	0.8333	$\langle 4,6;4,6\rangle$	3.2(1)
6.8541	9	0.8333	$\langle 3,4;8,8\rangle$	3.3
6.8541	9	0.8250	$\langle 4,4;5,8\rangle$	3.2(3)
6.8541	9	0.8095	$\langle 4,4;6,7\rangle$	3.2(3)
6.8541	9	0.8000	$\langle 5,5;5,5\rangle$	3.2(4)

many non–edge-homogeneous tessellations with exponential growth that are vertex-homogeneous or face-homogeneous and that grow more slowly. For example, the method of Lemma 6.2 in [11] can be applied to the unique 3-valent, face-homogeneous tessellation with vertex-sequence $(6,6,7)$; its growth rate is $\frac{1}{4}\big(1+\sqrt{13}+\sqrt{2\sqrt{13}-2}\big) \approx 1.722$. (Uniqueness follows from [14], p. 613.)

Notation. Thanks to Proposition 2.8, we are entitled to the notational convenience of writing $\mu(p,q,k,\ell)$, when we mean $\mu(T)$, where T is the (unique) tessellation with edge-symbol $\langle p,q;k,\ell\rangle$. For $t \geq 4$, let $\mathscr{T}(t)$ denote the set of tessellations T for which $t = \big(\frac{p+q}{2}-2\big)\big(\frac{k+\ell}{2}-2\big)$. Let $M(t)$ and $m(t)$ denote the greatest and the least value, respectively, of $\mu(T)$ for $T \in \mathscr{T}(t)$.

The data suggest some sort of inverse correlation between t-value (and hence growth rate) and the value of μ. We formulate this in terms of the following nonlinear integer optimization problem.

PROBLEM 5.1. *For fixed* $t \geq 4$, *assume that the tessellation* $T \in \mathscr{T}(t)$ *has edge-symbol* $\langle p,q;k,\ell\rangle$.

$$\begin{aligned} \textit{Maximize and minimize}: \quad & \mu(p,q,k,\ell) \quad \textit{subject to} \\ & p,q,k,\ell \geq 3 \\ & \textit{and} \\ & \left(\frac{p+q}{2}-2\right)\left(\frac{k+\ell}{2}-2\right) = t. \end{aligned}$$

The following theorem shows that $M(t)$ is strictly decreasing.

THEOREM 5.2. *For each* $t \geq 6$,

$$M(t) = \mu(3, t+1, 4, 4) = \frac{5}{6} + \frac{1}{t+1},$$

and the second largest value of μ *on* $\mathscr{T}(t)$ *is*

$$\mu(4, t, 4, 4) = \frac{3}{4} + \frac{1}{t}.$$

Proof. Assume that $t \geq 6$. Let $M_1 = M$, and let $M_2(t)$ denote the second largest value of μ on $\mathscr{T}(t)$. Clearly $\mathscr{T}(t)$ contains tessellations with edge-symbol $\langle 3, t+1; 4, 4\rangle$ and with edge-symbol $\langle 4, t; 4, 4\rangle$, and these tessellations yield the values of μ given in the statement of this theorem.

Without loss of generality, we may assume that $p \leq q$, $k \leq \ell$, and, by duality, $p \leq k$. If $p, q, k, \ell \geq 5$, then $\mu(p, q, k, \ell) < \frac{5}{6} < \mu(3, t+1, 4, 4) \leq M_1(t)$. If $p, q, k, \ell \geq 6$, then $\mu(p, q, k, \ell) < \frac{3}{4} < \mu(4, t, 4, 4) \leq M_2(t)$. Hence we need to consider only those edge-symbols with p equal to 3, 4, or 5.

Suppose that $p = 3$. By Proposition 2.8, T has edge-symbol $\langle 3, q; k, k\rangle$, where either $q = 3$ and k is odd or q is arbitrary and k is even. In the former instance, we must have $k = t + 2$ and $\mu(T) = \frac{2}{3} + \frac{2}{t+2}$. The assumption that $\mu(T) > \mu(4, t, 4, 4)$ leads to the quadratic inequality $0 > t^2 - 10t + 24$, whose only real integer solution is $t = 5$, contrary to assumption.

Now suppose that k is even. Then $t = \frac{1}{2}(q-1)(k-2)$. If $k = 4$, then $q = t + 1$, and $\mu(T) = M_1(t)$ as claimed. If $k = 6$, we obtain exactly the same contradiction as in the previous paragraph. If $k = 8$, then $q = \frac{1}{3}(t+3)$, and the assumption that $\mu(3, q, 8, 8) > \frac{3}{4} + \frac{1}{t}$ leads to the inequality $0 > t^2 - 9t + 18$, which implies that $t = 6$ and hence $q = 3$, in which case $\langle 3, 3; 8, 8\rangle$ ties for second place in $\mathscr{T}(6)$ (see Table 1). Finally, if $k \geq 10$, then $\mu(3, q, k, k) \leq \frac{8}{15} + \frac{1}{q}$. If this quantity is greater than $\frac{3}{4} + \frac{1}{t}$, then we must have $q = 4$, which leads to the same quadratic inequality as in the subcase of $k = 8$ but, in this instance, to a contradiction.

Suppose that $p = 4$. We may assume that $q \geq 5$ or $k \geq 5$. First suppose that q is odd. Then $k = \ell \geq 6$ and is even. If T has edge-symbol $\langle 4, 5; 6, 6\rangle$, then $T \in \mathscr{T}(10)$ and $\mu(4, 5, 6, 6) = 0.78\overline{3} < 0.85 = \mu(4, 10, 4, 4)$. If $q \geq 7$ or $k \geq 8$, then one easily checks that $\mu(p, q, k, \ell) < 3/4$.

Now suppose that $q = 4$. First consider the tessellation T with edge-symbol $\langle 4, 4; 6, \ell\rangle$. Then $T \in \mathscr{T}(\ell + 2)$ and $\mu(T) = \frac{2}{3} + \frac{1}{\ell}$. But if $\mu(4, \ell+2; 4, 4) \leq \mu(T)$, then $0 \geq \ell^2 + 2\ell - 24$, contrary to the assumption that $\ell \geq 6$. Hence $7 \leq k \leq \ell$. If T has edge-symbol $\langle 4, 4; 7, 7\rangle$, then $\mu(T) \approx 0.7857$, but $\mu(4, 10; 4, 4\rangle = 0.85$, as noted in the previous paragraph. Hence $8 \leq k \leq \ell$, and so $\mu(4, 4; k, \ell) \leq \frac{3}{4}$.

Hence $q \geq 6$. If $k \geq 6$, then $\mu(4, q; k, \ell) \leq \frac{3}{4}$, and so we need consider only the tessellation T with edge-symbol $\langle 4, 6; 4, 6\rangle$, which belongs to $\mathscr{T}(9)$. But then $\mu(T) = 0.8\overline{3}$, while $\mu(4, 9, 4, 4) = 0.86\overline{1}$.

Finally suppose that $p = 5$. Since 5 is odd, we have the same two possibilities as in the case of $p = 3$. If T has edge-symbol $\langle 5, 5; k, k\rangle$, then by our initial assumptions, $k \geq 5$. If $k = 5$, then $T \in \mathscr{T}(9)$ and $\mu(T) = 0.8 < \frac{3}{4} + \frac{1}{9} = \mu(4, 9, 4, 4)$. If $k \geq 6$, then clearly $\mu(T) < \frac{3}{4}$. Hence suppose that T has edge-symbol $\langle 5, q; k, k\rangle$, where $q \geq 6$. This forces k to be even, and so $k \geq 6$. But then $\mu(T) \leq 0.7$, completing the proof. ∎

A glance at Table 1 shows that the theorem holds for $M_1(5)$ but fails for $M_2(5)$. But what about the least values $m(t)$? Treating p, q, k, and ℓ as continuous variables

and applying the method of Lagrange multipliers (see any standard advanced calculus text), we find that, for fixed $t \geq 4$, a minimum of $\mu(p,q,k,\ell) = 4/(2+\sqrt{t})$ occurs when all four parameters equal $2+\sqrt{t}$. Thus when t is a perfect square, a tessellation of minimum μ-value $m(t)$ is found in $\mathscr{T}(t)$ by this formula. It turns out, however, that $m(t)$ does not decrease monotonically. As a counterexample, $m(12) = 0.73\overline{3}$ is realized by $\langle 5,5;6,6\rangle$. But $m(13) = 0.736\overline{1}$ is realized by $\langle 4,4;8,9\rangle$.

6. Initial conditions and enumeration. In this final section we demonstrate how the transition matrices constructed in section 4, in conjunction with elementary enumeration methods, can be used to obtain the ordinary generating function for the number of faces (or vertices or edges) in each corona for any Bilinski diagram of any edge-homogeneous tessellation with any given root. Although, for any given value of t, the individual parameters in an edge-symbol have no effect asymptotically as one computes the growth rate, these parameters do determine the initial conditions of the recurrence system and hence determine the concrete numbers that we are about to compute.

As mentioned earlier, the vector $\mathbf{v}_1$ is determined by the choice of the root of the Bilinski diagram. Suppose, for example, that the tessellation in question belongs to $\mathscr{M}_{4,4}$ and all valences and covalences are even (as in Case 1 of Lemma 4.2). If the root is a p-valent vertex, then

$$\mathbf{v}_1 = [0, p/2, p/2, 0, 0, 0]^t;$$

if the root is an edge, then

$$\mathbf{v}_1 = \left[1, \frac{p-2}{2}, \frac{q-2}{2}, 1, \frac{p-2}{2}, \frac{q-2}{2}\right]^t;$$

if the root is a k-covalent face, then

$$\mathbf{v}_1 = \left[0, \frac{k(p-2)}{4}, \frac{k(q-2)}{4}, k, \frac{k(p-4)}{4}, \frac{k(q-4)}{4}\right]^t. \tag{6.1}$$

Determination of the initial vector $\mathbf{v}_1$ in the other cases of the edge-symbol is a straightforward exercise.

We now demonstrate how to count the vertices in U_n for $n \geq 2$. This, of course, in a concentric Bilinski diagram equals the number of edges in the circuit $\langle U_n\rangle$ induced by U_n. Let us define the weighted row vector $\mathbf{w} = [w_1, \ldots, w_m]$, where w_i denotes the number of edges in $\langle U_n\rangle$ incident with a face in F_n of the ith face-type. Thus we have the dot product

$$|U_n| = \mathbf{w} \cdot \mathbf{v}_n^t, \qquad (n \geq 2). \tag{6.2}$$

For example, if the edge-homogeneous tessellation with edge-symbol $\langle p,q;k,\ell\rangle$ belongs to $\mathscr{M}_{4,4}$ and all parameters are even, then

$$\mathbf{w} = [k-3,\ k-2,\ k-2,\ \ell-3,\ \ell-2,\ \ell-2]. \tag{6.3}$$

We conclude this article by applying the foregoing computations to an example.

Example. Consider the tessellation T with edge-symbol $\langle 4,6;4,6\rangle$. As we are in Case 1 of Lemma 4.2, we have $t = 9$ and $\gamma(T) = g(9) = \frac{1}{2}\big(7+3\sqrt{5}\big) \approx 6.854$. The

transition matrix for T is

$$A = \begin{bmatrix} 0 & 0 & 0 & 3 & 4 & 4 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1/2 & 1 & 2 & 3/2 & 2 & 2 \\ 1 & 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 \\ 1/2 & 1 & 1 & 5/2 & 3 & 4 \end{bmatrix}.$$

We choose the root of our Bilinski diagram of T to be a 4-covalent face. By (6.1),

$$\mathbf{v}_1 = [0, 2, 4, 4, 0, 2]^t.$$

From Theorem 3.1, the closed form of the ordinary generating function φ for the sequence $\{|F_n| : n \geq 0\}$ is

$$\varphi(x) = 1 + \frac{12x}{x^2 - 7x + 1}.$$

The two roots of the denominator are $\lambda_1 = \frac{1}{2}\left(7 + 3\sqrt{5}\right) = \gamma(T)$ and $\lambda_2 = 1/\lambda_1 = \frac{1}{2}\left(7 - 3\sqrt{5}\right)$. Elementary algebra leads to

$$\varphi(x) = 1 + \frac{4}{15}\sum_{n=0}^{\infty}\left[(2\lambda_1 - 7)\lambda_1^n + (2\lambda_2 - 7)\lambda_2^n\right]x^n.$$

Thus

$$|F_n| = \begin{cases} 1 & \text{if } n = 0, \\ \frac{4}{15}\left[(2\lambda_1 - 7)\lambda_1^n + (2\lambda_2 - 7)\lambda_2^n\right] & \text{if } n \geq 1. \end{cases}$$

The sequence $\{|F_n| : n \geq 0\}$ begins with $1, 12, 84, 576, \ldots$.

Equation (6.3) yields the weighted vector $\mathbf{w} = [1, 2, 2, 3, 4, 4]$, from which we compute by (6.2) the sequence $\{|U_n| : n \geq 0\}$, which begins with $4, 32, 220, 1512, \ldots$.

REFERENCES

[1] S. Bilinski, *Homogene mreže ravnine*, Rad Jugoslav. Akad. Znan. Umjet., 271 (1948), pp. 145–255.

[2] S. Bilinski, *Homogene Netze der Ebene*, Bull. Internat. Acad. Yougoslave. Cl. Sci. Math. Phys. Tech. (N.S.), 2 (1949), pp. 63–111.

[3] C.P. Bonnington, W. Imrich, and M.E. Watkins, *Separating rays in locally finite, planar graphs*, Discrete Math., 145 (1995), pp. 61–72.

[4] J.A. Bruce, *Bilinski Diagrams and Geodesics in* 1*-Ended Planar Maps*, Doctoral dissertation, Syracuse University, Syracuse, NY, 2002.

[5] J.A. Bruce and M.E. Watkins, *Concentric Bilinski diagrams*, Australas. J. Combin., 30 (2004), pp. 161–174.

[6] H.S.M. Coxeter, *Regular Polytopes*, 2nd ed., Macmillan, New York, 1963.

[7] J.E. Graver and M.E. Watkins, *Locally Finite, Planar, Edge-transitive Graphs*, Mem. Amer. Math. Soc. 126, American Mathematical Society, Providence, RI, 1997.

[8] B. Grünbaum and G.C. Shephard, *Edge-transitive planar graphs*, J. Graph Theory, 11 (1987), pp. 141–155.

[9] B. Grünbaum and G.C. Shephard, *Tilings and Patterns*, W.H. Freeman and Company, New York, 1987.

[10] W. Imrich, *On Whitney's theorem on the unique embeddability of* 3*-connected planar graphs*, in Recent Advances in Graph Theory, M. Fiedler, ed., Academia Praha, Prague, 1975, pp. 303–306.

[11] J.F. Moran, *The growth rate and balance of homogeneous tilings in the hyperbolic plane*, Discrete Math., 173 (1997), pp. 151–186.
[12] P. Niemeyer and M.E. Watkins, *Geodetic rays and fibers in one-ended planar graphs*, J. Combin. Theory Ser. B, 69 (1997), pp. 142–163.
[13] R. Sedgewick and P. Flajolet, *An Inroduction to the Analysis of Algorithms*, Addison-Wesley, Reading, MA, 1996.
[14] J. Šiagiová and M.E. Watkins, *Covalence sequences of planar vertex-homogeneous maps*, Discrete Math., 307 (2007), pp. 599–614.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 19–53

AVERAGE SPECTRA AND MINIMUM DISTANCES OF LOW-DENSITY PARITY-CHECK CODES OVER ABELIAN GROUPS*

GIACOMO COMO† AND FABIO FAGNANI‡

Abstract. Ensembles of regular low-density parity-check codes over any finite Abelian group G are studied. The nonzero entries of the parity matrix are randomly chosen, independently and uniformly, from an arbitrary label group of automorphisms of G. Precise combinatorial results are established for the exponential growth rate of their average type-enumerating functions with respect to the code-length N. Minimum Bhattacharyya-distance properties are analyzed when such codes are employed over a memoryless G-symmetric transmission channel. In particular, minimum distances are shown to grow linearly in N with probability one, and lower bounds are provided for the typical asymptotic normalized minimum distance. Finally, some numerical results are presented, indicating that the choice of the label group strongly affects the value of the typical minimum distance.

Key words. low-density parity-check codes, group codes, minimum distance, type-spectrum, Ramanujan sums

AMS subject classifications. 94B12, 94B65, 11T24

DOI. 10.1137/070686615

1. Introduction.

Low-density parity-check (LDPC) codes have received a huge amount of attention in the last years. It is indeed the family of high-performance codes for which the deepest theoretical insight has been achieved. Their definition is quite simple: they are those binary-linear codes which can be described as kernels of matrices over the binary field $\mathbb{Z}_2$ with a "small" number of nonzero elements. Since the pioneering work [19], two streams of research are easily recognizable in the literature on LDPC codes. On the one hand, structural properties of such codes have been investigated: distance-spectra, minimum distances, and also capacity estimations under maximum-likelihood (ML) decoding [28, 29, 25, 37, 25, 26, 9, 15, 33]. On the other hand, they have been studied coupled with the well-known iterative decoding schemes [34, 35, 42, 31, 43, 24, 36, 14].

The need to use transmission channels with higher spectral efficiency naturally leads one to consider nonbinary codes and nonbinary LDPC codes. A typical example is provided by the m-PSK Gaussian channel. This is a channel accepting as possible input any element in the set $m\text{-PSK} := \{e^{\frac{2\pi}{m}li} \mid 1 \leq l \leq m\}$, while the received output is obtained by adding a homogeneous, zero-mean, two-dimensional Gaussian variable. When m is an integer power of 2—a case which is particularly relevant in practice—in principle binary codes can be used for transmission over this channel. Using any fixed bijection $\lambda : \mathbb{Z}_2^r \to 2^r\text{-PSK}$, binary-linear codes can be mapped into codes on the alphabet 2^r-PSK. The problem with this type of code is that, if $r > 2$, for any possible choice of λ they will not possess many of the symmetry properties that binary-linear codes enjoy on binary symmetric channels: Voronoi regions will not be congruent, Euclidean distance profiles will depend on the reference codeword, and

*Received by the editors March 28, 2007; accepted for publication (in revised form) June 6, 2008; published electronically October 24, 2008.
http://www.siam.org/journals/sidma/23-1/68661.html
†Dipartimento di Matematica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy. Current address: Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 77 Mass. Ave., Cambridge, MA 02139 (giacomo@mit.edu).
‡Dipartimento di Matematica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy (fabio.fagnani@polito.it).

the uniform error property will not hold. As a consequence the theoretical analysis becomes quite hard and design-criteria optimization exceedingly complicated; in [22, 4, 5] an average-coset approach has been adopted in order to overcome this problem. Actually, for such an input set, a much better candidate group structure is provided by the cyclic group $\mathbb{Z}_m$. Indeed, if we consider the natural map $\lambda : \mathbb{Z}_m \to m$-PSK (with $\lambda(l) = e^{\frac{2\pi i l}{m}}$), any subgroup $\mathcal{C} \subseteq \mathbb{Z}_m^N$ yields, through the embedding λ, a code over m-PSK possessing congruent Voronoi regions and invariant distance profiles [18, 27]. These codes (as well as the subgroups they come from) are called $\mathbb{Z}_m$-codes. All of this construction can be generalized to a broader family of transmission channels exhibiting symmetry with respect to the action of a finite group G, which will be called G-symmetric channels, and to a family of codes with group structure which will be called G-codes.

$\mathbb{Z}_m$-codes have been widely studied in the past (see, for instance, [3]). A remarkable fact is that, since $\mathbb{Z}_m$ is a commutative ring, they can be represented, as in the binary case, as images or kernels of matrices with coefficients in $\mathbb{Z}_m$. In this paper we are particularly interested in kernel representations: given a matrix Φ in $\mathbb{Z}_m^{L\times N}$,

$$\mathcal{C} := \{\boldsymbol{x} \in \mathbb{Z}_m^N \mid \Phi\boldsymbol{x} = 0\}$$

is obviously a $\mathbb{Z}_m$-code. Regular LDPC $\mathbb{Z}_m$-codes can easily be constructed by considering syndrome matrices with a fixed amount of nonzero elements both on each row and on each column and, as in the binary case, randomly selecting nonzero positions through a random-permutation approach. An interesting difference with respect to the binary case is the way to choose the nonzero elements of Φ. In this paper we will consider many different possibilities. Among them, we consider the so-called *unlabelled ensemble*, where nonzero elements are all equal to 1, and the *uniformly labelled ensemble*, where nonzero elements are instead, each one independently, chosen to be any possible invertible element in the ring $\mathbb{Z}_m$ with uniform probability. We will see that the latter ensemble will outperform the former. Of course our results could be extended to irregular LDPC ensembles, where the fraction of rows and columns with different amounts of nonzero entries (degree profile) is fixed, although this extension will not be considered here. Nonbinary LDPC codes have been considered for binary-input channels as well (see [32], for instance). In this case, they allow us to introduce a new design parameter, the choice of the nonzero entries in the parity matrix, to be optimized jointly with the degree profile.

LDPC codes over nonbinary alphabets were already introduced and studied in Gallager's seminal work [19]. Precisely, Gallager considered regular ensembles of LDPC $\mathbb{Z}_m$-codes with all nonzero entries equal to 1 (unlabelled ensembles in our terminology); he studied their Hamming distance-spectra and provided bounds for their error probabilities under ML and suboptimal iterative decoding over some highly symmetric channels. More recently, after the rediscovery of Gallager codes in the 1990s, LDPC codes over nonbinary fields, both for binary and nonbinary channels, have received a considerable amount of attention by researchers. In [13], the authors show empirical evidence that, appropriately choosing the values of the nonzero entries in the parity-check matrix, LDPC codes over the Galois field $\mathbb{F}_{2^r}$ perform better than the corresponding binary LDPC codes when used over binary-input output-symmetric channels. LDPC codes over $\mathbb{F}_{2^r}$ for binary-input output-symmetric channels have also been studied in [32] following a density-evolution approach. The works [4, 5, 17] contain quite a complete theoretical analysis of LDPC codes over finite fields for nonbinary channels considering both ML and belief-propagation decoding. Average

type-spectra of regular LDPC ensembles over $\mathbb{Z}_p$ in the special case when p is prime, and more in general over $\mathbb{F}_{p^r}$, have been studied in [17, 4]. In this case the structural theory of binary LDPC codes generalizes in an almost straightforward way. In particular it has been shown, using expurgation techniques and results from [39], that average type-spectra provide lower bounds to the typical error exponent of these ensembles and that this exponent can be made arbitrarily close to the random-coding one by allowing the density of the parity matrix to grow while keeping the rate constant. Finally the recent works [8, 30, 38] investigate the possibility of using hybrid nonbinary LDPC codes over multiple groups.

However, in the case of algebraic structures which are not fields (e.g., $\mathbb{Z}_m$ with nonprime m), the available theoretical results are very few. In [4], average type-spectra of unlabelled ensembles of LDPC $\mathbb{Z}_m$-codes have also been studied in the case when m is not prime, but there are no results on minimum Euclidean distances. In the papers [40, 1, 44] the case when m is not prime has been considered but mainly from an iterative-decoding perspective. Computer simulations have been reported in [40, 44] showing that, when mapped over the m-PSK constellation, LDPC $\mathbb{Z}_m$-codes guarantee better performance than their binary counterparts.

When m is not prime, the lack of field structure implies that the theory of $\mathbb{Z}_m$-codes itself (with no restriction on the density of their kernel representation) is not as simple as in the linear case. This issue has been addressed in [10, 11], where the capacity achievable by $\mathbb{Z}_m$-codes (and more in general by Abelian group codes) over symmetric channels—called $\mathbb{Z}_m$-capacity—has been characterized in terms of the capacities of the channels obtained by restricting the input to all nontrivial subgroups of $\mathbb{Z}_m$ (see (2.5) in section 2.3). For the m-PSK constellation (when m is an integer power of a prime) it has been proved that $\mathbb{Z}_m$-codes achieve capacity, while this is no longer the case for other possible geometrically uniform constellations. Type-spectra and minimum distances of ensembles of $\mathbb{Z}_m$-codes have been studied in [12], where it has been shown that the typical $\mathbb{Z}_8$-code asymptotically achieves the Gilbert–Varshamov bound of the 8-PSK AWGN channel. The study of the properties of group-code ensembles gives insight into the theory of LDPC codes over groups, since it allows one to distinguish between the possible loss in performance due to the group structure and the one due to the sparseness of the syndrome representation.

In this paper we will study in detail average type-spectra and minimum Bhattacharyya-distances of regular LDPC ensembles over any finite Abelian group G, in which the nonzero entries of the parity-check operator are randomly sampled, independently and uniformly, from an arbitrary group F of automorphisms of G (briefly F-labelled ensembles), generalizing all of the results in [19, 13, 17, 4]. This extension passes through the use of mathematical tools which do not show up in the binary case: group characters, arithmetic concepts (Möbius inversion formula, Ramanujan sums), combinatorial techniques (Cayley graphs), and convex-analytical techniques.

As a first result, we will find exact expressions in terms of combinatorial formulas for the average type-spectra of regular F-labelled ensembles of LDPC codes over G; see Theorem 3.5. For the unlabelled ensemble of LDPC codes over $\mathbb{Z}_m$, we will show that our results for average type-spectra coincide with those obtained in [19, 4], while for LDPC codes over finite fields the results of [13, 17, 4] will be recovered. Theorem 3.5 is instead completely original, to the best of our knowledge, for the uniformly labelled ensemble of LDPC codes over $\mathbb{Z}_m$, for which the average type-spectrum has an elegant expression in terms of Ramanujan sums. Coupling this analysis with an ad hoc analysis for the low-weight average type-enumerating functions, we will finally propose upper bounds to the probability distribution of the minimum Bhattacharyya distance

[6]. This will allow us to show that minimum distances grow linearly in N with probability one (see Theorems 5.1 and 5.2): in the coding terminology this means that such codes are asymptotically good with probability one. More precisely, we obtain almost sure lower bounds on the asymptotic normalized minimum distance of the LDPC ensembles. These bounds are defined as the solution of $(|G|-1)$-dimensional optimization problems. Proving the tightness of these bounds would require second-moment estimations for the type-enumerating functions and is a problem left for future research. However, concentration results available in the literature for the Hamming distance-spectra of regular ensembles of binary LDPC codes (see [33]) make us optimistic about the tightness of our bounds for regular ensembles of LDPC G-codes as well. Finally, we will present some numerical results for the average distance-spectra showing how strongly the choice of the label group F affects the value of the typical minimum distance. In particular, we will show that, for the 8-PSK AWGN channel, the distance properties of the uniformly labelled ensemble of LDPC $\mathbb{Z}_8$-codes are significantly better than those of the unlabelled ensemble. This is confirmed by Monte Carlo simulations of these codes which we have run, and it agrees with some of the simulation results reported in [4].

The remainder of this paper is organized as follows. Section 2 is devoted to a formal introduction of all of the fundamental concepts: G-symmetric channels and the associated Bhattacharyya distance, Abelian group codes and their capacity, and LDPC code ensembles over Abelian groups. In section 3 we study the average type-enumerating functions of these ensembles, and we determine their exact growth rate, namely the so-called average type-spectrum: the main result is Theorem 3.5. Section 4 is a technical one devoted to a detailed probabilistic analysis of low-weight codewords: the main result is Theorem 4.6. Using the results of sections 3 and 4 we are able to prove, in section 5, a probabilistic lower bound on the growth of minimum Euclidean distances for the LDPC ensembles when the block-length N goes to infinity; see Theorems 5.1 and 5.2. Finally, in section 6 we report some numerical simulations showing that the uniformly labelled ensemble of LDPC $\mathbb{Z}_8$-codes definitely outperforms the unlabelled one on the 8-PSK AWGN channel, and we draw some final conclusions. An appendix completes the paper, containing some of the most technical proofs and a technical lemma on semicontinuous functions.

2. The coding setting.

2.1. Notation. Throughout the paper $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, $\mathbb{C}$ will denote the usual number sets. With $\mathbb{R}_+$ ($\mathbb{Z}_+$) we will indicate the set of nonnegative reals (integers). If z is in $\mathbb{C}$, then z^* is its conjugate. The functions log and exp are to be considered with respect to a fixed base $a > 1$. Conventionally, $\inf(\emptyset) = +\infty$, $\sup(\emptyset) = -\infty$. For any subset $B \subseteq A$, $\overline{B} := A \setminus B$ will denote the complement of B in A, while $\mathbb{1}_B : A \to \{0,1\}$ will denote the indicator function of B, defined by $\mathbb{1}_B(a) = 1$ if a belongs to B and $\mathbb{1}_B(a) = 0$ otherwise. For a finite set A, $L^2(A)$ will denote the vector space of all $\mathbb{C}$-valued functions on A, equipped with the Hermitian form $\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \sum_{a \in A} \boldsymbol{f}(a)\boldsymbol{g}(a)^*$. For a function $\boldsymbol{f}$ in $L^2(A)$ we shall indicate by $\mathrm{supp}(\boldsymbol{f}) := \big\{a \in A \big| \, \boldsymbol{f}(a) \neq 0\big\}$ the support of $\boldsymbol{f}$. Given $\boldsymbol{f}, \boldsymbol{g} \in L^2(A)$, $\boldsymbol{f} \cdot \boldsymbol{g} \in L^2(A)$ will denote their pointwise product, while we define $\boldsymbol{f}^{\boldsymbol{g}} := \prod_{a \in \mathrm{supp}(\boldsymbol{f})} \boldsymbol{f}(a)^{\boldsymbol{g}(a)}$. We consider the simplex $\mathcal{P}(A)$ of probability measures on A, $\mathcal{P}(A) := \{\boldsymbol{\theta} : A \to \mathbb{R}_+ | \, \sum_a \boldsymbol{\theta}(a) = 1\}$. Given a subset $B \subseteq A$, we shall use the notation $\boldsymbol{\theta}(B) := \sum_{b \in B} \boldsymbol{\theta}(b)$. For a in A, δ_a in $\mathcal{P}(A)$ will be the probability measure concentrated on a. The entropy function $\mathrm{H} : \mathcal{P}(A) :\to \mathbb{R}$ and the Kullback–Leiber distance $D(\cdot||\cdot) : \mathcal{P}(A) \times \mathcal{P}(A) \to [0, +\infty]$

are defined, respectively, by

$$\mathrm{H}(\boldsymbol{\theta}) := -\sum_{a\in\mathrm{supp}(\boldsymbol{\theta})} \boldsymbol{\theta}(a)\log\boldsymbol{\theta}(a), \qquad D\left(\boldsymbol{\theta}||\boldsymbol{\theta}'\right) := \sum_{a\in\mathrm{supp}(\boldsymbol{\theta})} \boldsymbol{\theta}(a)\log\frac{\boldsymbol{\theta}(a)}{\boldsymbol{\theta}'(a)}.$$

Given $\boldsymbol{x}\in A^N$, its *A-type* (or empirical frequency) is the probability measure $\boldsymbol{\theta}_A(\boldsymbol{x})\in\mathcal{P}(A)$ given by $[\boldsymbol{\theta}_A(\boldsymbol{x})](a) = \frac{1}{N}|\{1\le i\le N : x_i = a\}|$. Define the set of types of all N-tuples by $\mathcal{P}_N(A) := \boldsymbol{\theta}_A(A^N)$, and let $\mathcal{P}_{\mathbb{N}}(A) := \bigcup_N \mathcal{P}_N(A)$ be the set of all A-types. The number of A-types $|\mathcal{P}_N(A)| = \binom{N+|A|-1}{|A|-1}$ is a quantity growing polynomially fast in N. Instead, the set of N-tuples of a given type $\boldsymbol{\theta}$, denoted by

$$A^N_{\boldsymbol{\theta}} := \left\{\boldsymbol{x}\in A^N \text{ such that (s.t.) } \boldsymbol{\theta}_A(\boldsymbol{x}) = \boldsymbol{\theta}\right\},$$

has cardinality growing exponentially fast with N. More precisely, for $\boldsymbol{\theta}\in\mathcal{P}_{\mathbb{N}}(A)$, consider the set $\mathcal{N}_{\boldsymbol{\theta}} := \{N : N\boldsymbol{\theta}(a)\in\mathbb{N}\ \forall a\in A\}$ which is infinite since $|A|\mathbb{N}\subseteq\mathcal{N}_{\boldsymbol{\theta}}$. Then, for every $N\in\mathcal{N}_{\boldsymbol{\theta}}$, we have $\left|A^N_{\boldsymbol{\theta}}\right| = \binom{N}{N\boldsymbol{\theta}} := N!/\prod_a (N\boldsymbol{\theta}(a))!$, and Stirling's formula implies that

$$\left|A^N_{\boldsymbol{\theta}}\right| \le \exp(N\,\mathrm{H}(\boldsymbol{\theta})), \qquad \lim_{N\in\mathcal{N}_{\boldsymbol{\theta}}} \frac{1}{N}\log\left|A^N_{\boldsymbol{\theta}}\right| = \mathrm{H}(\boldsymbol{\theta}). \tag{2.1}$$

2.2. Symmetric channels. A memoryless channel (MC) is described by a finite input set $\mathcal{X}$, an output set consisting of a σ-finite measure space $\mathcal{Y} = (Y,\mathcal{B},\nu)$, and a family of transition probability densities $P(\cdot|x)$ on $\mathcal{Y}$ indexed by the possible inputs x in $\mathcal{X}$. Such a channel will be denoted by $(\mathcal{X},\mathcal{Y},P)$. In the applications there are essentially two possibilities: either Y is finite and ν is simply the counting measure (and in this case $P(\cdot|x)$ are simply probabilities on $\mathcal{Y}$), or $\mathcal{Y}$ is an n-dimensional Euclidean space and ν is the corresponding Lebesgue measure. Keeping this more abstract formalism will allow us to cover both cases at once.

We now recall the concept of a group action. Given a finite group G with identity 1_G and a (finite) set A, we say that G acts on A if, for every g in G, it is defined as a map from A to A denoted by $a\mapsto ga$, such that $1_G a = a$ for all a in A and $h(ga) = (hg)a$ for all h,g in G and a in A. The action of G over A is said to be (simply) transitive if for every $a,b\in A$ there exists one (and only one) element g of G such that $ga = b$. If the action is simply transitive, G and A are clearly in bijection: $g\mapsto ga_0$, where a_0 is some fixed reference element in A.

Given a σ-finite measure space $\mathcal{Y} = (Y,\mathcal{B},\nu)$, we say that the group G acts isometrically on $\mathcal{Y}$ if it is defined as an action of G on Y consisting of measurable bijections such that

$$\nu(gA) = \nu(A) \qquad \forall A\in\mathcal{B},\ \forall g\in G. \tag{2.2}$$

Notice that in the case, when Y is a finite set, (2.2) is trivially always verified so that in this case all actions are isometric. Instead, in the case when $Y = \mathbb{R}^n$, (2.2) is a real restriction and is verified if the maps $y\mapsto gy$ are isometries of $\mathbb{R}^n$.

DEFINITION 2.1. *An MC $(\mathcal{X},\mathcal{Y},P)$ is said to be* G-symmetric *if the following hold:*

(a) *there exists a simply transitive action of G on $\mathcal{X}$;*
(b) *there exists an isometric action of G on $\mathcal{Y}$;*
(c) *$P(y|x) = P(gy|gx)$ for every $g\in G$, every $x\in\mathcal{X}$, and ν-almost every $y\in\mathcal{Y}$.*

It follows from (a) that the input $\mathcal{X}$ of a G-symmetric MC and the group G are in bijection: we will often tend to identify them. In this paper we will exclusively consider the case when G is a finite Abelian group. We present a few fundamental examples.

Example 1 (binary-input output-symmetric channels). Consider the case when $G \simeq \mathbb{Z}_2$. $\mathbb{Z}_2$-symmetric channels are known in the coding literature as binary-input output-symmetric (BIOS) channels. Typical examples are the binary symmetric channel (BSC) and the binary erasure channel (BEC). By considering r consecutive uses of a BIOS channel $(\mathcal{X}, \mathcal{Y}, P)$, one obtains a $\mathbb{Z}_2^r$-symmetric MC with input set $\mathcal{X}^r$, output space $\mathcal{Y}^r$, and product transition probabilities $P(\boldsymbol{y}|\boldsymbol{x}) := \prod_{1\le k\le r} P(y_k|x_k)$.

Example 2 (m-ary symmetric channel). Consider a finite set $\mathcal{X}$ of cardinality $m \ge 2$ and some $\varepsilon \in [0,1]$. The *m-ary symmetric channel* is described by the triple $(\mathcal{X}, \mathcal{X}, P)$, where $P(y|x) = 1-\varepsilon$ if $y = x$ and $P(y|x) = \varepsilon/(m-1)$ otherwise. This channel returns the transmitted input symbol x as output with probability $1-\varepsilon$, while with probability ε a wrong symbol is received, uniformly distributed over the set $\mathcal{X} \setminus \{x\}$. The special case $m = 2$ corresponds to the BSC. The m-ary symmetric channel was considered by Gallager [19, sect. 5] to evaluate the performance of nonbinary LDPC codes. It exhibits the highest possible level of symmetry. Indeed, it is G-symmetric for every group G of order $|G| = m$. To see this, it is sufficient to observe that every group acts simply and transitively on itself. Notice that whenever $m = p^r$ for some prime p and positive integer r, the group G can be chosen to be $\mathbb{Z}_p^r$, which is compatible with the structure of the Galois field $\mathbb{F}_{p^r}$.

Example 3 (geometrically uniform AWGN channels). An n-dimensional constellation is a finite subset $S \subset \mathbb{R}^n$ spanning $\mathbb{R}^n$. We denote with $\mathrm{Iso}(S)$ its symmetry group, i.e., the group of those isometries of $\mathbb{R}^n$ mapping S into S itself. A constellation S is said to be geometrically uniform (GU) if there exists a subgroup G of $\mathrm{Iso}(S)$ whose action on S is simply transitive. Such a G is called a generating group for S: for every $s \in S$ the mapping $\lambda_s : G \to S$ defined by $\lambda_s : g \in G \mapsto gs \in S$ is a bijection called isometric labeling.

Given a GU constellation $S \subset \mathbb{R}^n$ with generating group G, define the *S-AWGN channel* as the n-dimensional unquantized AWGN channel with input set S, output $\mathbb{R}^n$ with the usual Borel–Lebesgue measure structure, and transition probability densities given by $P(y|x) = N(y-x)$, where $N(x) = (2\pi\sigma^2)^{-n/2} e^{-||x||^2/2\sigma^2}$ is the density of an n-dimensional diagonal Gaussian random variable. Now let S' be another GU constellation such that $S \subseteq S'$ and G is isomorphic to a subgroup of $\mathrm{Iso}(S')$. Let us introduce the quantization map over the Voronoi regions of S' $q : \mathbb{R}^n \to S'$, $q(x) = \mathrm{argmin}_{s\in S'} ||x-s||$ (resolving nonuniqueness cases by assigning to $q(x)$ a value arbitrarily chosen from the set of minima). We define the *(S,S')-AWGN channel* as the MC obtained by applying q to the output of the S-AWGN channel. Note that the special case $S = S'$ coincides with the so-called hard decoding rule. It is easy to see that both the S-AWGN channel and the (S,S')-AWGN channel are G-symmetric.

The simplest example of a GU constellation is the so-called one-dimensional antipodal constellation $\{-1, 1\}$, admitting $\mathbb{Z}_2$ as a generating group. Another example is given by m orthogonal equal-energy signals: in this case the symmetry group coincides with the permutation group S_m, and any group of order m is a generating group. A two-dimensional example is the m-PSK constellation already introduced in section 1. Notice that the symmetry group of the m-PSK is isomorphic to D_m, the dihedral group with $2m$ elements. m-PSK always admits cyclic generating group $\mathbb{Z}_m$. When m is even, the m-PSK also admits a generating group isomorphic to $D_{m/2}$, which is non-Abelian for all $m \ge 6$. Notice that the only cases when m-PSK has a generating group admitting Galois field structure are when m is prime or $m = 4$.

In fact, when $m = 2^r$ with $r \geq 3$ or $m = p^r$ with $p \geq 3$ prime and $r \geq 2$, $\mathbb{Z}_p^r$ is not isomorphic to any subgroup of D_m and thus cannot be a generating group for m-PSK.

Consider an MC $(\mathcal{X}, \mathcal{Y}, P)$ and two input elements x, x' in $\mathcal{X}$. The Schwarz inequality gives

$$0 \leq \int_{\mathcal{Y}} \sqrt{P(y|x)P(y|x')}\mathrm{d}\nu(y) \leq \int_{\mathcal{Y}} P(y|x)\mathrm{d}\nu(y) \int_{\mathcal{Y}} P(y|x')\mathrm{d}\nu(y) = 1.$$

Moreover, the first of the previous inequalities holds as an equality iff $P(\,\cdot\,|x)P(\,\cdot\,|x') = 0$ ν-almost everywhere. Instead, the second inequality is equality iff $P(\cdot|x) = P(\cdot|x')$ ν-almost everywhere, which means that actually x and x' have indistinguishable outputs. Throughout this paper we will assume that $0 < \int_{\mathcal{Y}} \sqrt{P(y|x)P(y|x')}\mathrm{d}\nu(y) < 1$ for every $x \neq x'$. While there is no loss of generality in the latter part of this assumption, the former excludes from our analysis the class of channels whose 0-error capacity is strictly positive. To any MC we can associate a function

$$\boldsymbol{\Delta} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ , \qquad \boldsymbol{\Delta}(x, x') := -\log \int_{\mathcal{Y}} \sqrt{P(y|x)P(y|x')}\mathrm{d}\nu(y).$$

This function is usually called the *Bhattacharyya distance* (or simply $\boldsymbol{\Delta}$-distance) of the channel. $\boldsymbol{\Delta}$ is symmetric: $\boldsymbol{\Delta}(x, x') = \boldsymbol{\Delta}(x', x)$; moreover, $\boldsymbol{\Delta}(x, x') = 0$ iff $x = x'$. The Bhattacharyya distance can be extended to direct products in a natural way. Given $\boldsymbol{x}, \boldsymbol{x}'$ in $\mathcal{X}^N$, we put $\boldsymbol{\Delta}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^N \boldsymbol{\Delta}(x_i, x_i')$. The *minimum* $\boldsymbol{\Delta}$*-distance* of a code $\mathcal{C} \subseteq \mathcal{X}^N$ is defined as

$$\mathrm{d}_{\min}(\mathcal{C}) := \min\{\boldsymbol{\Delta}(\boldsymbol{x}, \boldsymbol{x}') \;\big|\; \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{C},\ \boldsymbol{x} \neq \boldsymbol{x}'\}.$$

If the MC $(\mathcal{X}, \mathcal{Y}, P)$ is G-symmetric, it is easy to verify that $\boldsymbol{\Delta}(gx, gx') = \boldsymbol{\Delta}(x, x')$ for all x, x' in $\mathcal{X}$ and g in G. Identifying $\mathcal{X}$ with G as usual, we can introduce the so-called *Bhattacharyya weight*:

$$\boldsymbol{\delta} : G \to \mathbb{R}_+, \qquad \boldsymbol{\delta}(x) := \boldsymbol{\Delta}(x, 1_G), \qquad x \in G.$$

In this way we have $\boldsymbol{\Delta}(x, x') = \boldsymbol{\delta}(x^{-1}x')$.

In the case of a BIOS channel, we have that

$$\boldsymbol{\Delta}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{1 \leq i \leq N} \boldsymbol{\delta}(x_i - x_i') = \boldsymbol{\delta}(1)\,|\{1 \leq i \leq N :\, x_i \neq x_i'\}| \qquad \forall\, \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^N;$$

i.e., the $\boldsymbol{\Delta}$-distance is proportional to the Hamming distance (the number of different entries of two strings).

For the m-ary symmetric channel of Example 2 we obtain

$$\boldsymbol{\Delta}(\boldsymbol{x}, \boldsymbol{x}') = -\log\left(\varepsilon\tfrac{m-2}{m-1} + \sqrt{\tfrac{(1-\varepsilon)\varepsilon}{m-1}}\right)|\{1 \leq i \leq N \,:\, x_i \neq x_i'\}| \qquad \forall\, \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^N,$$

so that, once again, the $\boldsymbol{\Delta}$-distance is proportional to the Hamming distance.

Finally, for the S-AWGN channel of Example 3, by considering any isometric labeling $\lambda_s : G \to S$, we obtain

$$\begin{aligned} \boldsymbol{\Delta}(\boldsymbol{x}, \boldsymbol{x}') &= \sum_{k=1}^N -\log \int_{\mathbb{R}^n} \frac{e^{-\left(||\boldsymbol{y} - \lambda_s(x_k)||^2 + ||\boldsymbol{y} - \lambda_s(x_k')||^2\right)/4\sigma^2}}{(2\pi\sigma^2)^{n/2}}\mathrm{d}\boldsymbol{y} \\ &= \frac{\log e}{8\sigma^2} \sum_{k=1}^N ||\lambda_s(x_k) - \lambda_s(x_k')||^2 ; \end{aligned}$$

i.e., the Bhattacharyya distance is proportional to the squared Euclidean distance.

2.3. Group codes and type-enumerating functions. When transmitting over an MC which is symmetric according to Definition 2.1, a natural class of codes to be considered is that of group codes. A *G-code* of length N is any subgroup of the direct group product G^N. Group codes are generalizations of binary-linear codes (the latter correspond to the case $G \simeq \mathbb{Z}_2$). In fact, G-codes enjoy many of the properties of binary-linear codes. For instance, when a G-code $\mathcal{C}$ is employed on a G-symmetric MC, ML decision regions (Voronoi regions in the Gaussian case) are congruent, and then the error probability does not depend on the transmitted codeword: this is called the uniform error property [18].

For every G-code $\mathcal{C}$ of length N we now introduce some combinatorial quantities characterizing its performance. The *type-enumerating function* of a G-code $\mathcal{C}$ is defined as

$$W_{\mathcal{C}} : \mathcal{P}(G) \to \mathbb{Z}_+, \qquad W_{\mathcal{C}}(\boldsymbol{\theta}) := \sum_{\boldsymbol{x} \in G^N_{\boldsymbol{\theta}}} \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) \qquad \forall\, \boldsymbol{\theta} \in \mathcal{P}(G),$$

where $G^N_{\boldsymbol{\theta}}$ is the set of N-tuples of type $\boldsymbol{\theta}$. Notice that since $\mathcal{C}$ is a subgroup of G^N, 1_{G^N} is always a codeword so that $W_{\mathcal{C}}(\delta_{1_G}) = 1$.

Assume we have fixed a G-symmetric MC $(\mathcal{X}, \mathcal{Y}, P)$, and let $\boldsymbol{\delta}$ be its associated Bhattacharyya weight. The minimum $\boldsymbol{\Delta}$-distance of a G-code $\mathcal{C}$ of length N is a function of its type-enumerating function:

(2.3)
$$\mathrm{d}_{\min}(\mathcal{C}) = \min\{\boldsymbol{\delta}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{C} \setminus \{\mathbf{0}\}\} = N \inf\left\{\langle \boldsymbol{\delta}, \boldsymbol{\theta}\rangle \middle| \, \boldsymbol{\theta} \in \mathcal{P}(G) \setminus \{\delta_0\} : W_{\mathcal{C}}(\boldsymbol{\theta}) > 0\right\}.$$

Type-enumerating functions and minimum Bhattacharyya distances play an important role in the estimation of the ML decoding error probability of G-codes over G-symmetric MCs. For instance, the so-called union-Bhattacharyya bound, for the error probability of a G-code $\mathcal{C}$ of length N, can be written in the form

$$p_e(\mathcal{C}) \leq \sum_{\boldsymbol{\theta} \in \mathcal{P}(G)} W_{\mathcal{C}}(\boldsymbol{\theta}) \exp\left(-N\langle \boldsymbol{\delta}, \boldsymbol{\theta}\rangle\right). \tag{2.4}$$

Bounds tighter than (2.4) can be obtained for the error probability of G-codes over G-symmetric channels based on variations of the Gallager bound [20, 39].

Observe that both (2.3) and (2.4) do not generally hold when a G-code is employed on an MC which is not G-symmetric. While this is not an issue for the highly symmetric channels considered in Example 2, it does matter for the symmetric channels introduced in Example 3. As a concrete example, one may think of the 8-PSK Gaussian channel: in this case, while both (2.3) and (2.4) are true for $\mathbb{Z}_8$-codes, for a $\mathbb{Z}_2^3$-code $\mathcal{C}$, and a fortiori for a $\mathbb{F}_8$-linear code, neither (2.3) nor (2.4) holds. In fact, the type-enumerating function of a $\mathbb{Z}_2^3$-code is not sufficient for characterizing its performance on the 8-PSK Gaussian channel. In order to overcome this problem, an average coset ensemble approach needs to be used [22, 4, 5].

It is a well-known result in information theory [20] that binary-linear codes allow one to achieve the capacity of any BIOS channels. More in general, linear codes over the Galois field $\mathbb{F}_{p^r}$ are known to achieve the capacity of any $\mathbb{Z}_p^r$-symmetric channel [16]. A similar result was conjectured in [27] for G-codes on G-symmetric MCs. In [11], the capacity C_G achievable by G-codes on G-symmetric MCs has been characterized for any finite Abelian group G. When G is cyclic of order

$m = p_1^{r_1^m} p_2^{r_2^m} \dots p_s^{r_s^m}$, for distinct primes $p_1, \dots, p_s$, it has been shown that

$$C_{\mathbb{Z}_m} = \max_{\boldsymbol{\alpha} \in \mathcal{P}(\{1,\dots,s\})} \min_{l | m, l>1} \frac{C_{p^s}}{\sum_{1 \le j \le s} \boldsymbol{\alpha}(j) \frac{r_j^l}{r_j^m}} \le C, \tag{2.5}$$

where C_l denotes the Shannon capacity of the $\mathbb{Z}_l$-symmetric channel obtained by restricting the input of the original channel to the subgroup $\frac{m}{l}\mathbb{Z}_m$. It has been shown in [11] that for a wide class of G-symmetric channels, including the p^r-PSK Gaussian channel (for prime p) both with quantized and unquantized output, G-capacity C_G and Shannon capacity C do coincide, while this is no longer the case for other G-symmetric channels.

2.4. LDPC codes over Abelian groups. For any finite Abelian group G, we now describe the ensembles of LDPC G-codes which will be considered in this paper. For every given degree pair (c, d) in $\mathbb{N}^2$, we consider the set of admissible block-lengths $\mathcal{N}_{(c,d)} := \{N \in \mathbb{N} \text{ s.t. } d \mid Nc\}$, and for every N in $\mathcal{N}_{(c,d)}$ we define $L = Nc/d$. Consider the c-repetition operator

$$\mathrm{Rep}_c^N : G^N \to G^{Nc}, \qquad (\mathrm{Rep}_c^N \boldsymbol{x})_i = x_{\lceil i/c \rceil}, \tag{2.6}$$

where $\lceil x \rceil$ denotes the lowest integer not below x, and the d-check summation operator

$$\mathrm{Sum}_d^N : G^{Nc} \to G^L, \qquad (\mathrm{Sum}_d^N \boldsymbol{x})_i = \sum_{k=i(d-1)+1}^{id} x_k. \tag{2.7}$$

Consider the group of permutations on Nc elements, S_{Nc}, and let Π'_N be a random variable uniformly distributed over S_{Nc}. Moreover, consider a subgroup F of $\mathrm{Aut}(G)$, the automorphism group of G, and let $(\Lambda_j)_{1 \le j \le Nc}$ be a family of independent random variables identically distributed uniformly on F, independent of Π'_N. Define the random diagonal automorphism $\Pi''_N \in \mathrm{Aut}(G^{Nc})$ by $(\Pi''_N \boldsymbol{x})_j := \Lambda_j x_j$ for $1 \le j \le Nc$. Finally, for every $N \in \mathcal{N}_{(c,d)}$ define the random syndrome homomorphism

$$\Phi_N : G^N \to G^L, \qquad \Phi_N := \mathrm{Sum}_d^N \, \Pi'_N \Pi''_N \, \mathrm{Rep}_c^N \tag{2.8}$$

and the associated random G-code $\mathcal{C}_N := \ker \Phi_N$. This is called the (c, d)-*regular* F-*labelled ensemble*. F will be called the *label group*. The two extreme cases $F = \{1\}$ and $F = \mathrm{Aut}(G)$ will be referred to, respectively, as the *unlabelled* and the *uniformly labelled* (c, d)-regular ensembles.

The reason for considering only automorphisms as possible labels, avoiding the use of noninvertible labels, is clarified by the following proposition. For any group H, we denote the set of endomorphisms of H by $\mathrm{End}(H)$.

PROPOSITION 2.2. *Assume that, for all $N \in \mathcal{N}_{(c,d)}$, $\Phi_N : G^N \to G^L$ is defined as in* (2.8) *with Π'_N uniformly distributed over S_{Nc} and $\Pi''_N \in \mathrm{End}(G^{Nc})$ is defined by $(\Pi''_N \boldsymbol{x})_j := \Lambda_j x_j$ for $1 \le j \le Nc$, where (Λ_j) are independently and identically distributed according to some probability distribution $\boldsymbol{\mu} \in \mathcal{P}(\mathrm{End}(G))$ such that $\mathrm{supp}(\boldsymbol{\mu}) \not\subseteq \mathrm{Aut}(G)$. Then, for all $k \in G \setminus \{0\}$ such that $\Lambda k = 0$ for some $\Lambda \in \mathrm{supp}(\boldsymbol{\mu})$*

$$\mathbb{P}(\mathrm{d}_{\min}(\ker \Phi_N) \le \boldsymbol{\delta}(k)) \ge 1 - (1 - \boldsymbol{\mu}(\Lambda)^c)^N \overset{N \to \infty}{\longrightarrow} 1.$$

Proof. Consider $\Lambda \in \mathrm{supp}(\boldsymbol{\mu}) \setminus \mathrm{Aut}(G)$, and $k \in \ker \Lambda \setminus \{0\}$. For $1 \le s \le N$, let $e_s^k \in G^N$ be the N-tuple with all-zero entries but the sth one, which is equal to k. If

$\Lambda_j = \Lambda$ for all $(s-1)c+1 \le j \le sc$, then $\Pi''_N \operatorname{Rep}_c^N e_s^k = \mathbf{0}$, so that $\Phi_N e_s^k = \mathbf{0}$, and $\mathrm{d}_{\min}(\ker \Phi_N) \le \boldsymbol{\delta}(k)$. Since the events

$$E_s^N := \bigcap_{(s-1)c+1 \le j \le sc} \{\Lambda_j = \Lambda\}$$

are independent for $1 \le s \le N$ and all have probability $1 - \boldsymbol{\mu}(\Lambda)^c$, it follows that

$$\mathbb{P}(\mathrm{d}_{\min}(\ker \Phi_N) \le \boldsymbol{\delta}(k)) \ge \mathbb{P}\left(\bigcup_{1 \le s \le N} E_s^N \right) = 1 - \left(1 - \mathbb{P}\left(E_s^N\right)\right)^N = \left(1 - \boldsymbol{\mu}(\Lambda)^c\right)^N . \quad \square$$

We wish to underline the fact that the proof of Proposition 2.2 strongly relied on the independence assumption we made for the labels Λ_j. Indeed, by introducing proper dependence structures for the random labels which allow us to avoid certain configurations, it is possible to consider ensembles of LDPC G-codes with noninvertible labels as well. This possibility will not be considered in the present paper but will be explored in a future work.

As LDPC G-codes are special G-codes admitting sparse kernel representation, they suffer from all of the limitations in performance of G-codes. In particular, the capacity they can achieve on a G-symmetric channel is upper bounded by the G-capacity of that channel. This explains why the authors of [4] had to restrict themselves to prime values of m while studying LDPC $\mathbb{Z}_m$-codes, albeit the average type-spectra they obtained for the unlabelled ensemble did not need such an assumption. In fact, they noticed that for nonprime m "expurgation is impossible" and LDPC $\mathbb{Z}_m$-codes result "bounded away from the random-coding spectrum." The same restriction to prime values of m (or more in general to groups G admitting Galois field structure) was required both in [4] and [17] in order to study the uniformly labelled ensemble.

In this paper regular ensembles of F-labelled LDPC G-codes will be studied for any finite Abelian group G. In particular we will find estimations for their average type-enumerating functions $\overline{W_{\mathcal{C}_N}(\boldsymbol{\theta})}$ and explicit combinatorial formulas for their average type-spectra defined as the limit of $N^{-1} \log \overline{W_{\mathcal{C}_N}(\boldsymbol{\theta})}$. Coupling this analysis with an ad hoc analysis of the type-enumerator functions for small weight codewords, we will finally propose upper bounds to the repartition function of the minimum normalized distance $\frac{1}{N} \mathrm{d}_{\min}(\mathcal{C}_N)$. This will allow us to show that, if $c > 2$, minimum distances grow linearly in N with high probability. We will also show that the typical minimum distance (more precisely the lower bound on it—conjectured to be tight—provided by the average type-spectra) of the uniformly labelled LDPC ensemble is significantly larger than the typical minimum distance of the corresponding unlabelled ensemble.

In [10] it was claimed that, for any m, the (c, d)-regular ensemble allows one to achieve a nonzero capacity over any $\mathbb{Z}_m$-symmetric channel, and that this capacity can be made arbitrarily close to the $\mathbb{Z}_m$-capacity of the channel, if the parameters (c, d) are allowed to grow. In fact, the same is true for the uniformly labelled ensembles as well; see section 6.2. This implies that LDPC $\mathbb{Z}_m$-codes allow one to achieve the Shannon capacity of a $\mathbb{Z}_m$-symmetric channel whenever $\mathbb{Z}_m$-codes do. While explicit proofs of these facts will not be given here due to the lack of space, they can be obtained from the combinatorial results of sections 3 and 4 using standard upper-bounding techniques for the average error probability of group codes [20, 39]. Similar reasonings can be made for minimum distances and error exponents of LDPC codes. In particular, minimum Bhattacharyya distances of $\mathbb{Z}_m$-codes have been studied in [12].

3. Average type-spectra of LDPC G-codes. In this section we first present some considerations on semidirect-product group actions. Then in section 3.2 we introduce LDPC codes in a slightly more general setting, and we show how regular F-labelled ensembles of LDPC G-codes introduced in section 2.4 can be cast in this framework. In section 3.3 we prove the main result, Theorem 3.5, characterizing the average type-spectra of regular F-labelled ensembles. Finally, in section 3.4 we show how previous results in the literature can be recovered as particular cases of Theorem 3.5, and we provide an explicit formula for the average type-spectrum of the uniformly labelled ensemble over the cyclic group, which is instead an original result.

3.1. Group actions. We recall here some basic facts about semidirect group actions; the reader is referred to the standard textbook [23] for further details. Assume that a group F acts on a set A. A subset $B \subseteq A$ is said to be F-invariant if $fb \in B$ for every $b \in B$ and $f \in F$. Clearly, if B is F-invariant, F acts on B as well. For every a in A, the relative orbit $Fa := \{b \in A \text{ s.t. } b = fa \text{ for some } f \in F\}$ is F-invariant and its action on it is transitive. The set of orbits is denoted by A/F and called the quotient of A by the action of F. There is a canonical surjection $\pi_F : A \to A/F$ which associates an element a with the orbit it belongs to. Given $a \in A$, we define its stabilizer as $\mathrm{Stab}_F(a) := \{f \in F \text{ s.t. } fa = a\}$. The well-known class formula gives $|F| = |Fa| \cdot |\mathrm{Stab}_F(a)|$.

If A and B are sets and the group F acts on A, a map $\phi : A \to B$ is said to be F-invariant if $\phi(fa) = \phi(a)$ for every $a \in A$ and $f \in F$. As an example, the canonical surjection $\pi_F : A \to A/F$ is an F-invariant map. Suppose we have an F-invariant map $\phi : A \to B$; then it is immediate to see that we can define a map $\tilde{\phi} : A/F \to B$ such that $\phi = \tilde{\phi} \circ \pi_F$. Notice that if it happens that ϕ is onto and moreover $\phi(a) = \phi(a')$ iff $Fa = Fa'$, then the map $\tilde{\phi}$ is a bijection, and thus A/F and B are in one-to-one correspondence. We will often use this fact in order to characterize quotient spaces.

We now introduce an example which will play a fundamental role in our future derivations. Given any set A, the permutation group S_N acts naturally on A^N: given $\boldsymbol{a} \in A^N$ and σ in S_N, we define $\sigma\boldsymbol{a} \in A^N$ by $(\sigma\boldsymbol{a})_j = \boldsymbol{a}_{\sigma^{-1}(j)}$. Orbits can easily be described using types. Given $\boldsymbol{a}, \boldsymbol{b} \in A^N$, it is immediate to see that

$$\exists\, \sigma \in S_N \ : \ \sigma\boldsymbol{a} = \boldsymbol{b} \ \Leftrightarrow \ \boldsymbol{\theta}_A(\boldsymbol{a}) = \boldsymbol{\theta}_A(\boldsymbol{b}).$$

This says that the subsets $A^N_{\boldsymbol{\theta}}$ of type-$\boldsymbol{\theta}$ N-tuples are exactly the orbits for the action of the permutation group S_N on A^N, and we have a natural bijection $A^N/S_N \simeq \mathcal{P}_N(A)$ (obtained through the mapping $\boldsymbol{a} \mapsto \boldsymbol{\theta}_A(\boldsymbol{a})$).

Now suppose we are given an action of a group F on the set A. This extends to a componentwise action of F^N on A^N with the orbit set $A^N/F^N \simeq (A/F)^N$. We would like to combine this action with the action of the permutation group on A^N, and the way to do this is as follows: we consider the semidirect product

$$S_N \ltimes F^N, \qquad (\sigma_1, \boldsymbol{g}_1)(\sigma_2, \boldsymbol{g}_2) = (\sigma_1\sigma_2, (\sigma_2^{-1}\boldsymbol{g}_1)\boldsymbol{g}_2)$$

and the action on A^N given by $(\sigma, \boldsymbol{g})\boldsymbol{a} = \sigma(\boldsymbol{g}\boldsymbol{a})$.

We now want to characterize the set of orbits of this semidirect action. Notice that the map $\pi_F : A \to A/F$ induces a natural map $\pi_F^\sharp : \mathcal{P}(A) \to \mathcal{P}(A/F)$, where $[\pi_F^\sharp\boldsymbol{\theta}](Fa) = \sum_{b \in Fa} \boldsymbol{\theta}(b)$. It is easy to see that the following diagram commutes:

$$\begin{array}{ccc} A^N & \stackrel{\pi_{F^N}}{\to} & (A/F)^N \\ \downarrow \boldsymbol{\theta}_A & & \downarrow \boldsymbol{\theta}_{A/F} \\ \mathcal{P}_N(A) & \stackrel{\pi_F^\sharp}{\to} & \mathcal{P}_N(A/F) \end{array} \tag{3.1}$$

(i.e., $\boldsymbol{\theta}_{A/F} \circ \pi_{F^N} = \pi_F^\sharp \circ \boldsymbol{\theta}_A$).

In what follows we will use the notation $\boldsymbol{\nu}_{A,F} = \boldsymbol{\theta}_{A/F} \circ \pi_{F^N}$ and call $\boldsymbol{\nu}_{A,F}(\boldsymbol{a})$ the (A,F)-*type* of $\boldsymbol{a}$. The (A,F)-type is exactly what is needed to describe orbits with respect to the action of the semidirect group $S_N \ltimes F^N$. Indeed, it is immediate to check that $\mathcal{P}_N(A/F)$ is in bijection with the quotient $A^N/(S_N \ltimes F^N)$: given $\boldsymbol{a}, \boldsymbol{b} \in A^N$, we have that

$$\exists (\sigma, \boldsymbol{g}) \in S_N \ltimes F^N \text{ s.t. } (\sigma, \boldsymbol{g})\boldsymbol{a} = \boldsymbol{b} \Leftrightarrow \boldsymbol{\nu}_{A,F}(\boldsymbol{a}) = \boldsymbol{\nu}_{A,F}(\boldsymbol{b}).$$

If $\boldsymbol{\nu} \in P_N(A/F)$, we will use the notation $A^N_{\boldsymbol{\nu}} := \{\boldsymbol{a} \in A^N \mid \boldsymbol{\nu}_{A,F}(\boldsymbol{a}) = \boldsymbol{\nu}\}$. Using the fact that $\boldsymbol{\nu}_{A,F} = \boldsymbol{\theta}_{A/F} \circ \pi_{F^N}$ we obtain that

$$|A^N_{\boldsymbol{\nu}}| = \tbinom{N}{N\boldsymbol{\nu}} \prod_{\alpha \in A/F} |\pi_F^{-1}(\alpha)|^{N\boldsymbol{\nu}(\alpha)}. \tag{3.2}$$

Now define $\mathcal{O}^N_{\boldsymbol{\nu}} := \{\boldsymbol{\theta} \in \mathcal{P}_N(A) \text{ s.t. } \pi_F^\sharp(\boldsymbol{\theta}) = \boldsymbol{\nu}\}$. For every given $\boldsymbol{\nu} \in \mathcal{P}(A/F)$, and N in $\mathbb{N}$, we have

$$A^N_{\boldsymbol{\nu}} = \bigcup_{\boldsymbol{\theta} \in \mathcal{O}^N_{\boldsymbol{\nu}}} A^N_{\boldsymbol{\theta}}, \tag{3.3}$$

the union being disjoint. Notice that we also have $\left|\mathcal{O}^N_{\boldsymbol{\nu}}\right| = \prod_{\alpha \in A/F} |\pi_F^{-1}(\alpha)|^{N\boldsymbol{\nu}(\alpha)}$.

3.2. A general framework for LDPC ensembles over Abelian groups. Fix an infinite subset $\mathcal{N} \subseteq \mathbb{N}$, a group U, two sequences of finite Abelian groups $Z^{(N)}$ and $Y^{(N)}$ (with $N \in \mathcal{N}$), and two sequences of homomorphisms

$$\Xi_o^N : U^N \to Z^{(N)}, \qquad \Xi_i^N : Z^{(N)} \to Y^{(N)}.$$

Consider, moreover, a sequence I_N of subgroups of $\mathrm{Aut}(Z^{(N)})$, and assume that the actions of I_N on $Z^{(N)}$ satisfy the following property: there exists a fixed finite set A and a sequence of invariant maps $\Theta_N : Z^{(N)} \to \mathcal{P}(A)$ such that $\boldsymbol{x}$, $\boldsymbol{y} \in Z^{(N)}$ are in the same orbit iff $\Theta_N(\boldsymbol{x}) = \Theta_N(\boldsymbol{y})$. In this way the quotient space $Z^{(N)}/I_N$ can be naturally identified with the image of Θ_N inside $\mathcal{P}(A)$.

Now let Π_N be a sequence of random variables uniformly distributed over I_N. For every $N \in \mathcal{N}$ define

$$\Phi_N := \Xi_i^N \Pi_N \Xi_o^N. \tag{3.4}$$

The triple (Ξ_o^N, Ξ_i^N, I_N) is called an *interconnected ensemble*, while $(\ker \Phi_N)$ will be the *random code sequence* associated with the ensemble. The set A will be called the *interconnection type alphabet* of the ensemble.

Now consider the type-enumerating function $W_N(\boldsymbol{\theta})$ for the ensemble. By taking the expectation with respect to our probability space, we get

$$\overline{W_N(\boldsymbol{\theta})} = \mathbb{E}\Big[\sum_{\boldsymbol{x} \in U^N_{\boldsymbol{\theta}}} \mathbb{1}_{\{\mathbf{0}\}}(\Phi_N \boldsymbol{x})\Big] = \sum_{\boldsymbol{x} \in U^N_{\boldsymbol{\theta}}} \mathbb{P}(\Phi_N \boldsymbol{x} = \mathbf{0}). \tag{3.5}$$

Put $Z^{(N)}_{\boldsymbol{\nu}} := \Theta_N^{-1}(\boldsymbol{\nu})$, and define the following sets: for every $\boldsymbol{\nu} \in \mathcal{P}(A)$, $\boldsymbol{\theta} \in \mathcal{P}(U)$

$$Z^{i,N}_{\boldsymbol{\nu}} := \left\{\boldsymbol{w} \in Z^{(N)}_{\boldsymbol{\nu}} \mid \Xi_i^N \boldsymbol{w} = \mathbf{0}\right\}, \qquad U^{o,N}_{\boldsymbol{\theta},\boldsymbol{\nu}} := \{\boldsymbol{x} \in U^N \mid \boldsymbol{\theta}_U(\boldsymbol{x}) = \boldsymbol{\theta},\ \Theta_N(\Xi_o^N \boldsymbol{x}) = \boldsymbol{\nu}\}. \tag{3.6}$$

We have the following simple result.

PROPOSITION 3.1. *For every* $\boldsymbol{\theta}$ *in* $\mathcal{P}_N(U)$

$$\overline{W_N(\boldsymbol{\theta})} = \sum_{\boldsymbol{v}\in\mathcal{P}(A)} \frac{|U^{o,N}_{\boldsymbol{\theta},\boldsymbol{v}}||Z^{i,N}_{\boldsymbol{v}}|}{|Z^{(N)}_{\boldsymbol{v}}|}. \tag{3.7}$$

Proof. If $\boldsymbol{x} \in U^{o,N}_{\boldsymbol{\theta},\boldsymbol{v}}$, using the fact that I_N acts transitively on $Z^{(N)}_{\boldsymbol{v}}$ and the class formula, we obtain

$$\mathbb{P}(\Phi_N \boldsymbol{x} = \mathbf{0}) = \mathbb{P}(\Pi_N \Xi^N_o \boldsymbol{x} \in Z^{i,N}_{\boldsymbol{v}}) = \frac{|Z^{i,N}_{\boldsymbol{v}}||\mathrm{Stab}_{I_N}(\Xi^N_o(\boldsymbol{x}))|}{|I_N|} = \frac{|Z^{i,N}_{\boldsymbol{v}}|}{|Z^{(N)}_{\boldsymbol{v}}|}.$$

Now using (3.5), (3.7) follows immediately. ∎

We now frame the LDPC ensembles introduced in section 2 into this more general setting. We use the notation introduced in section 2.4. Given $(c,d) \in \mathbb{N}^2$ and $N \in \mathcal{N}_{(c,d)}$, consider $L = Nc/d$. Take $U = G$, $Z^{(N)} = G^{Nc}$, $Y^{(N)} = G^L$. Also, take $\Xi^N_o = \mathrm{Rep}^N_c$, $\Xi^N_i = \mathrm{Sum}^N_d$, $I_N = S_{Nc} \ltimes F^{Nc}$. The ensemble $(\mathrm{Rep}^N_c, \mathrm{Sum}^N_d, S_{Nc} \ltimes F^{Nc})$ is the (c,d)-regular F-labelled ensemble. The type alphabet in this case is simply $A = G/F$.

Irregular ensembles can be framed into this setting by simply modifying the repetition and the sum operators. Also other interesting cases can be obtained by considering the interconnections among the inner and outer encoder done through some vector structured channels and allowing only independent permutations on the various channels. Finally, hybrid nonbinary LDPC codes can be considered in this framework by replacing the product group U^N with the product of copies of different Abelian groups $U^N_1 \times \cdots \times U^N_k$.

However, we will now focus on the evaluation of the type-spectra of the regular F-labelled LDPC G-code ensembles. This will be done in the following subsection by explicitly calculating the three terms entering in the formula (3.7).

3.3. The average type-spectrum of the (c,d)-regular F-labelled ensemble. In order to prove the main result of this section we will use some generating function techniques. For a finite set A, consider the ring of complex-coefficient multivariable polynomials (briefly multinomials) $\mathbb{C}[A]$. Given $p \in \mathbb{C}[A]$ and $\boldsymbol{k} \in \mathbb{Z}^A_+$, we denote by $\lfloor p(\boldsymbol{z}) \rfloor_{\boldsymbol{k}}$ the coefficient of the term $\boldsymbol{z}^{\boldsymbol{k}}$ in $p(\boldsymbol{z})$, i.e., $p(\boldsymbol{z}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^A_+} \lfloor p(\boldsymbol{z}) \rfloor_{\boldsymbol{k}} \boldsymbol{z}^{\boldsymbol{k}}$. In particular, we will consider type-enumerating multinomials, i.e., homogeneous-degree multinomials of the form $p(\boldsymbol{z}) = \sum_{\boldsymbol{\theta}\in\mathcal{P}_N(A)} \lfloor p(\boldsymbol{z}) \rfloor_{N\boldsymbol{\theta}} \boldsymbol{z}^{N\boldsymbol{\theta}}$, where each coefficient $\lfloor p(\boldsymbol{z}) \rfloor_{N\boldsymbol{\theta}}$ equals the number of N-tuples $\boldsymbol{a} \in A^N$ of A-type $\boldsymbol{\theta}$, satisfying certain properties. The easiest case is provided by the multinomial $(\sum_{a\in A} z_a)^N = \sum_{\boldsymbol{\theta}\in\mathcal{P}_N(A)} \binom{N}{N\boldsymbol{\theta}} \boldsymbol{z}^{N\boldsymbol{\theta}}$, simply enumerating the N-tuples of different A-types. The following result, proved in [9], characterizes the asymptotic growth rate of the coefficients of powers of enumerating multinomials.

THEOREM 3.2. *Let A be a finite set and $p(\boldsymbol{z}) \in \mathbb{R}_+[A]$ be a homogeneous-degree, nonnegative, real-coefficient multinomial. For all $\boldsymbol{\theta} \in \mathcal{P}_{\mathbb{N}}(A)$ and $\boldsymbol{z} \in \mathcal{P}(A)$ such that* $\mathrm{supp}(\boldsymbol{z}) = \mathrm{supp}(\boldsymbol{\theta})$*, we have*

$$\lfloor p(\boldsymbol{z})^N \rfloor_{N\boldsymbol{\theta}} \le \frac{p(\boldsymbol{z})^N}{\boldsymbol{z}^{N\boldsymbol{\theta}}}, \qquad \lim_{N\in\mathcal{N}_{\boldsymbol{\theta}}} \frac{1}{N} \log \lfloor p(\boldsymbol{z})^N \rfloor_{N\boldsymbol{\theta}} = \inf_{\substack{\boldsymbol{z}\in\mathcal{P}(A):\\ \mathrm{supp}(\boldsymbol{z})=\mathrm{supp}(\boldsymbol{\theta})}} \log \frac{p(\boldsymbol{z})}{\boldsymbol{z}^{\boldsymbol{\theta}}}. \tag{3.8}$$

Moreover, the left-hand side of (3.8) *is a concave (and thus upper semicontinuous)* $[-\infty,+\infty)$*-valued function on* $\mathcal{P}(A)$.

Observe that, by considering $p(\boldsymbol{z}) = \sum_a z_a$, (2.1) can be deduced from Theorem 3.2.

The first type-enumerating multinomial which we will need in our derivations is the one enumerating the 0-sum d-tuples over a finite Abelian group G:

$$\beta_d(\boldsymbol{z}) \in \mathbb{C}[z_g,\, g \in G], \qquad \beta_d(\boldsymbol{z}) := \sum_{g_1,\dots,g_d} \mathbb{1}_{\{0\}}\left(\sum_{k=1}^{d} g_k\right) \prod_{1\le k\le d} z_{g_k}.$$

By introducing the group $\hat{G}$ of characters of G, i.e., homomorphisms of G in the multiplicative group $\mathbb{C}^*$ of nonzero complex numbers, it is possible to find an explicit expression for $\beta_d(\boldsymbol{z})$ as stated in the following lemma.

LEMMA 3.3. *For every finite Abelian group* G *and* $d \in \mathbb{N}$

$$\beta_d(\boldsymbol{z}) = \frac{1}{|G|} \sum_{\chi\in\hat{G}} \left(\sum_{g\in G} z_g \chi(g)\right)^d.$$

Proof. The inversion formula for the discrete Fourier transform (see [41, p. 168]) $f(g) = \frac{1}{|G|}\sum_\chi \langle f,\chi\rangle \chi(g)$, applied to $f = \delta_0 \in L^2(G)$, gives $\frac{1}{G}\sum_\chi \chi(g) = \mathbb{1}_{\{0\}}(g)$. Then

$$\begin{aligned}
\beta_d(\boldsymbol{z}) &= \sum_{g_1,\dots,g_d} \mathbb{1}_{\{0\}}\Big(\sum_{1\le k\le d} g_k\Big) \prod_{1\le k\le d} z_{g_k} \\
&= \sum_{g_1,\dots,g_d} \tfrac{1}{|G|}\sum_\chi \chi\Big(\sum_{1\le k\le d} g_k\Big) \prod_{1\le k\le d} z_{g_k} \\
&= \tfrac{1}{|G|}\sum_\chi \sum_{g_1,\dots,g_d} \prod_{1\le k\le d} \chi\left(g_k\right) z_{g_k} \\
&= \tfrac{1}{|G|}\sum_\chi \Big(\sum_g z_g \chi(g)\Big)^d. \qquad \square
\end{aligned}$$

Recall that, given any subgroup F of $\mathrm{Aut}(G)$ and a degree pair (c,d) in $\mathbb{N}^2$, the (c,d)-regular F-labelled ensemble of LDPC G-codes is described by the triple $\big(\mathrm{Rep}_c^N, \mathrm{Sum}_d^N, S_{Nc} \ltimes F^{Nc}\big)$. Let $\pi_F : G \to G/F$ be the canonical projection on the quotient and $\pi_F^\sharp : \mathcal{P}(G) \to \mathcal{P}(G/F)$ be the associated action on probabilities. Also, define

$$\boldsymbol{\varphi} : G/F \to \mathbb{N}, \qquad \boldsymbol{\varphi}(q) = \left|\pi_F^{-1}(q)\right| \tag{3.9}$$

to be the map giving the cardinalities of the orbits of G under the action of F.

Consider some admissible block-length N in $\mathcal{N}_{(c,d)}$. Formula (3.2) shows that $|Z_{\boldsymbol{v}}^{(N)}| = \binom{Nc}{Nc\boldsymbol{v}} \boldsymbol{\varphi}^{Nc\boldsymbol{v}}$ for every $\boldsymbol{v} \in \mathcal{P}_{Nc}(G/F)$. Moreover, in this case $|U_{\boldsymbol{\theta},\boldsymbol{v}}^{o,N}| = \binom{N}{N\boldsymbol{\theta}} \mathbb{1}_{\{\pi_F^\sharp\boldsymbol{\theta}\}}(\boldsymbol{v})$. Substituting into (3.7), and defining $\boldsymbol{v} := \pi_F^\sharp\boldsymbol{\theta}$, we obtain

$$\overline{W_N(\boldsymbol{\theta})} = \binom{N}{N\boldsymbol{\theta}}\binom{Nc}{Nc\boldsymbol{v}}^{-1} \boldsymbol{\varphi}^{-Nc\boldsymbol{v}} |Z_{\boldsymbol{v}}^{i,N}|. \tag{3.10}$$

It remains to evaluate the enumerating weights $|Z_{\boldsymbol{\upsilon}}^{i,N}|$ relative to the check summation operator. In order to do that, we introduce the multinomial

$$\alpha_{F,d}(\boldsymbol{t}) \in \mathbb{C}[t_q,\, q \in G/F], \qquad \alpha_{F,d}(\boldsymbol{t}) := \frac{1}{|G|}\sum_{\chi\in\hat{G}}\left(\sum_{q\in G/F}\frac{1}{\boldsymbol{\varphi}(q)}\sum_{g\in q}\chi(g)t_q\right)^d \tag{3.11}$$

and present the following result, stating that the Lth power of $\alpha_{F,d}(\boldsymbol{t})$ is the type-enumerating multinomial of the normalized weights $|Z_{\boldsymbol{\upsilon}}^{i,N}|/\boldsymbol{\varphi}^{Nc\boldsymbol{\upsilon}}$.

LEMMA 3.4. *For every* $N \in \mathcal{N}_{(c,d)}$

$$\sum_{\boldsymbol{\upsilon}\in\mathcal{P}_{Nc}(G/F)}\frac{|Z_{\boldsymbol{\upsilon}}^{i,N}|}{\boldsymbol{\varphi}^{Nc\boldsymbol{\upsilon}}}\boldsymbol{t}^{Nc\boldsymbol{\upsilon}} = \left(\alpha_{F,d}(\boldsymbol{t})\right)^L. \tag{3.12}$$

Proof. First, consider the type-enumerating multinomial $B(\boldsymbol{z}) \in \mathbb{C}[z_g,\, g \in G]$ for the kernel of the inner homomorphism $\Xi_i^N = \mathrm{Sum}_d^N$. Since any $\boldsymbol{x}$ in G^{Nc} belongs to $\ker \mathrm{Sum}_d^N$ iff it is the concatenation of L 0-sum d-tuples, from Lemma 3.3 we have $B(\boldsymbol{z}) = \left(\beta_d(\boldsymbol{z})\right)^L$. Now consider the map

$$\Psi : \mathbb{C}[z_g,\, g \in G] \to \mathbb{C}[t_q,\ q \in G/F], \qquad \Psi : p(\boldsymbol{z}) \mapsto p(t_{\pi_F(g)},\, g \in G).$$

It follows from (3.3) that, for all $\boldsymbol{\upsilon}$ in $\mathcal{P}(G/F)$, we have

$$\frac{|Z_{\boldsymbol{\upsilon}}^{i,N}|}{\boldsymbol{\varphi}^{Nc\boldsymbol{\upsilon}}} = \sum_{\boldsymbol{\theta}\in\mathcal{O}_{\boldsymbol{\upsilon}}^{Nc}}\frac{\lfloor B(\boldsymbol{z})\rfloor_{Nc\boldsymbol{\theta}}}{\boldsymbol{\varphi}^{Nc\boldsymbol{\upsilon}}} = \sum_{\boldsymbol{\upsilon}\in\mathcal{P}_{Nc}(G/F)}\frac{\lfloor \Psi B\,(\boldsymbol{t})\rfloor_{Nc\boldsymbol{\upsilon}}}{\boldsymbol{\varphi}^{Nc\boldsymbol{\upsilon}}} = \sum_{\boldsymbol{\upsilon}\in\mathcal{P}_{Nc}(G/F)}\left\lfloor \Psi B\left(\frac{\boldsymbol{t}}{\boldsymbol{\varphi}}\right)\right\rfloor_{Nc\boldsymbol{\upsilon}}. \tag{3.13}$$

Thus, the claim follows by observing that $\Psi B\left(\boldsymbol{t}/\boldsymbol{\varphi}\right) = \left(\Psi\beta_d\left(\boldsymbol{t}/\boldsymbol{\varphi}\right)\right)^L = \alpha_{F,d}(\boldsymbol{t})^L$. ∎

We are now ready to prove the main result of this section, stating that the average type-spectrum of the (c,d)-regular F-labelled ensemble of LDPC G-codes is given by

$$\Gamma_{(F,c,d)}(\boldsymbol{\theta}) := \mathrm{H}(\boldsymbol{\theta}) + \frac{c}{d}\inf_{\substack{\boldsymbol{t}\in\mathcal{P}(G/F):\\ \mathrm{supp}(\boldsymbol{t})=\mathrm{supp}(\pi_F^\sharp\boldsymbol{\theta})}}\left\{\log\alpha_{F,d}(\boldsymbol{t}) + dD\big(\pi_F^\sharp\boldsymbol{\theta}||\boldsymbol{t}\big)\right\}. \tag{3.14}$$

From Theorem 3.2 it follows that the spectrum $\Gamma_{(F,c,d)}(\boldsymbol{\theta})$ is an upper semicontinuous function on the probability simplex $\mathcal{P}(G)$. Notice that, by choosing $\boldsymbol{t} = \pi_F^\sharp\boldsymbol{\theta}$, we immediately obtain the estimate

$$\Gamma_{(F,c,d)}(\boldsymbol{\theta}) \le \frac{c}{d}\log\alpha_{F,d}\left(\pi_F^\sharp\boldsymbol{\theta}\right) + \mathrm{H}(\boldsymbol{\theta}).$$

THEOREM 3.5. *For the* (c,d)*-regular* F*-labelled ensemble of LDPC* G*-codes*

$$\lim_{N\in\mathcal{N}_{\boldsymbol{\theta}}\cap\mathcal{N}_{(c,d)}}\frac{1}{N}\log\overline{W_N(\boldsymbol{\theta})} = \Gamma_{(F,c,d)}(\boldsymbol{\theta}).$$

Proof. From (3.10), by recalling that $Nc = Ld$ and $\boldsymbol{\upsilon} = \pi_F^\sharp\boldsymbol{\theta}$, we get

$$\frac{1}{N}\log\overline{W_N(\boldsymbol{\theta})} = \frac{1}{N}\log\binom{N}{N\boldsymbol{\theta}} + \frac{c}{d}\frac{1}{L}\log\frac{|Z_{\boldsymbol{\upsilon}}^{i,N}|}{\binom{Ld}{Ld\boldsymbol{\upsilon}}\boldsymbol{\varphi}^{Ld\boldsymbol{\upsilon}}}.$$

From (2.1) we have $\lim \frac{1}{N} \log \binom{N}{N\boldsymbol{\theta}} = \mathrm{H}(\boldsymbol{\theta})$. Then we can first apply Lemma 3.4 and then Theorem 3.2 (notice that (3.12) with $L = 1$ implies that $\alpha_{F,d}(\boldsymbol{t})$ has nonnegative real coefficients and homogeneous degree), obtaining

$$\lim_N \frac{1}{L} \log \frac{|Z^{i,N}_{\boldsymbol{\upsilon}}|}{\binom{Ld}{Ld\boldsymbol{\upsilon}} \boldsymbol{\varphi}^{Ld\boldsymbol{\upsilon}}} = \lim_N \frac{1}{L} \log \frac{\left\lfloor \alpha_{F,d}(\boldsymbol{t})^L \right\rfloor_{Ld\boldsymbol{\upsilon}}}{\binom{Ld}{Ld\boldsymbol{\upsilon}} \boldsymbol{\varphi}^{Ld\boldsymbol{\upsilon}}}$$
$$= \inf_{\substack{\boldsymbol{t} \in \mathcal{P}(G/F): \\ \mathrm{supp}(\boldsymbol{t}) = \mathrm{supp}(\boldsymbol{\upsilon})}} \left\{ \log \frac{\alpha_{F,d}(\boldsymbol{t})}{\boldsymbol{t}^{d\boldsymbol{\upsilon}}} - d\,\mathrm{H}(\boldsymbol{\upsilon}) \right\}. \qquad \square$$

3.4. Special cases of Theorem 3.5. Now we particularize Theorem 3.5 to some important special cases, showing that all previously known results can be reobtained, while new interesting cases can be studied as well.

3.4.1. LDPC codes over Galois fields. Suppose $G \simeq \mathbb{Z}_p^r$ for some prime number p and positive integer r. First, let F coincide with the whole automorphism group $\mathrm{Aut}(\mathbb{Z}_p^r)$, which is isomorphic to the general linear group of $r \times r$ invertible matrices on $\mathbb{Z}_p$. In this case the probability that an N-tuple $\boldsymbol{x}$ in G^N belongs to the random LDPC code $\mathcal{C}_N = \ker\left(\mathrm{Sum}_d^N \, \Pi_N \, \mathrm{Rep}_c^N\right)$ depends only on the Hamming weight (i.e., number of nonzero entries) of $\boldsymbol{x}$. Indeed, it is easily seen that the action of $\mathrm{Aut}(\mathbb{Z}_p^r)$ on $\mathbb{Z}_p^r$ has only two orbits: one containing the zero element only and one containing all of the nonzero elements of $\mathbb{Z}_p^r$. Thus, the quotient space is $G/F = \{q_0, q_1\}$, with $\boldsymbol{\varphi}(q_0) = 1$, $\boldsymbol{\varphi}(q_1) = p^r - 1$. Moreover, since all nontrivial characters are orthogonal to the trivial one $\chi_0 \equiv 1$, it follows that $\sum_{g \in q_1} \chi(g) = -\chi(0) = -1$ for all $\chi \in \hat{G} \setminus \{\chi_0\}$. Then the average type-spectra of the (c,d)-regular $\mathrm{Aut}(\mathbb{Z}_p^r)$-labelled ensemble of LDPC $\mathbb{Z}_p^r$-codes are given by

(3.15)
$$\Gamma_{(\mathrm{Aut}(\mathbb{Z}_p^r),c,d)}(\boldsymbol{\theta}) = \mathrm{H}(\boldsymbol{\theta}) + \tfrac{c}{d} \inf_{t \in (0,1)} \left\{ \log\left(\tfrac{1}{p^r} + \tfrac{p^r-1}{p^r} \left(1 - \tfrac{p^r}{p^r-1} t\right)^d \right) + dD(\lambda||t) \right\},$$

where $\lambda := 1 - \boldsymbol{\theta}(0)$ and $D(\lambda||t) := \lambda \log \frac{\lambda}{t} + (1-\lambda) \log \frac{1-\lambda}{1-t}$.

Now consider the case $G \simeq \mathbb{Z}_p^r$ again, but now with label group $F \simeq \mathbb{F}_{p^r}^*$, the multiplicative group of nonzero elements of the Galois field $\mathbb{F}_{p^r}$. Observe that $\mathbb{F}_{p^r}^*$ can always be identified with a subgroup (proper if $r > 1$) of $\mathrm{Aut}(\mathbb{Z}_p^r)$. Nevertheless, the action of $\mathbb{F}_{p^r}^*$ on $\mathbb{Z}_p^r$ has the same two orbits as the action of the whole $\mathrm{Aut}(\mathbb{Z}_{p^r})$ on $\mathbb{Z}_p^r$. This shows that the (c,d)-regular $F_{p^r}^*$-labelled ensemble has the same average type-spectrum of the $\mathrm{Aut}(\mathbb{Z}_p^r)$-labelled ensemble, i.e.,

$$\Gamma_{(\mathbb{F}_{p^r}^*,c,d)}(\boldsymbol{\theta}) = \Gamma_{(\mathrm{Aut}(\mathbb{Z}_p^r),c,d)}(\boldsymbol{\theta}) \qquad \forall\, \boldsymbol{\theta} \in \mathcal{P}(\mathbb{Z}_p^r). \tag{3.16}$$

The expression (3.15) coincides with the spectrum of the $\mathbb{F}_{p^r}^*$-labelled ensemble obtained in [4, 17]. We observe that in [32] it was numerically observed that the density-evolution dynamical system [34] exhibits the same threshold value for the $\mathbb{F}_{p^r}^*$-labelled and the $\mathrm{Aut}(\mathbb{Z}_p^r)$-labelled ensembles over the BEC. Formula (3.16) shows that these ensembles have identical average type-spectra.

3.4.2. Unlabelled LDPC ensembles over cyclic groups. We now consider the case when $G \simeq \mathbb{Z}_m$ and $F = \{1\}$. In this case, the characters of $\mathbb{Z}_m$ are given by $\chi_k(h) := e^{\frac{2\pi}{m} hki}$ for $h, k \in \mathbb{Z}_m$, while, trivially, the quotient space $\mathbb{Z}_m/F$ coincides

with $\mathbb{Z}_m$ itself and $\boldsymbol{\varphi} \equiv 1$ (see (3.9)). It follows that

$$\alpha_{\{1\},d}(\boldsymbol{t}) = \beta_d(\boldsymbol{t}) = \tfrac{1}{m} \sum_{1 \le k \le m} \left(\sum_{1 \le h \le m} e^{\frac{2\pi}{m} hki} z_h \right)^d .$$

Then the average type-spectrum takes the following form:

(3.17)
$$\Gamma_{(\{1\},c,d)}(\boldsymbol{\theta}) = \mathrm{H}(\boldsymbol{\theta}) + \frac{c}{d} \inf_{\substack{\boldsymbol{z} \in \mathcal{P}(\mathbb{Z}_m) \\ \mathrm{supp}(\boldsymbol{z}) = \mathrm{supp}(\boldsymbol{\theta})}} \left\{ \log\left(\frac{1}{m} \sum_k \left(\sum_h e^{\frac{2\pi}{m} hki} z_h \right)^d \right) + dD\big(\boldsymbol{\theta} || \boldsymbol{z}\big) \right\}.$$

The above spectrum coincides with the one obtained in [4] (see also [19, p. 49]).

3.4.3. Uniformly labelled ensembles over cyclic groups. Finally, consider the case when $G \simeq \mathbb{Z}_m$ again, but this time with F isomorphic to $\mathbb{Z}_m^*$, the multiplicative group of units of $\mathbb{Z}_m$. Notice that $\mathbb{Z}_m^*$ acts by multiplication on the ring $\mathbb{Z}_m$. It is immediate to see that two $a, b \in \mathbb{Z}_m$ are in the same orbit with respect to this group action iff $(m, a) = (m, b)$, where (k, h) denotes the greatest common divisor of two naturals k and h. The quotient space $\mathbb{Z}_m / \mathbb{Z}_m^*$ can be identified with the set of divisors of m, $\mathbb{D}_m := \{l \in \mathbb{N} \text{ s.t. } l \mid m\}$. We have $|\mathbb{Z}_m^*| = \varphi(m)$, where $\varphi : \mathbb{N} \to \mathbb{N}$, $\varphi(n) = \big|\big\{m \in \mathbb{N} \text{ s.t. } m \le n,\, (n, m) = 1\big\}\big|$, is the Euler φ-function. The projection map is

$$\pi_{\mathbb{Z}_m^*} : \mathbb{Z}_m \to \mathbb{D}_m, \qquad \pi_{\mathbb{Z}_m^*}(a) = \frac{m}{(m, a)}.$$

Notice that, for every $l \in \mathbb{D}_m$, the orbit $\pi^{-1}_{\mathbb{Z}_m^*}(l)$ coincides with $\frac{m}{l}\mathbb{Z}_m^*$ and it is in bijection with $\mathbb{Z}_l^*$ through the map $h \mapsto \frac{m}{l} h$. Then $\boldsymbol{\varphi}(l) = |\pi^{-1}_{\mathbb{Z}_m^*}(l)| = |\mathbb{Z}_l^*| = \varphi(l)$.

In order to evaluate the average-type spectra of the (c, d)-regular $\mathbb{Z}_m^*$-labelled ensemble of LDPC $\mathbb{Z}_m$-codes, it is convenient to introduce the so-called Ramanujan sums

$$r_l(k) := \sum_{j \in \mathbb{Z}_l^*} e^{\frac{2\pi}{l} jki}, \qquad l, k \in \mathbb{N}.$$

The Ramanujan sums are well known in number theory and can be explicitly evaluated in terms of both the Euler φ-function and Möbius function:

$$\mu : \mathbb{N} \to \mathbb{Z}, \qquad \mu(n) = \begin{cases} 1 & \text{if} \quad n = 1, \\ 0 & \text{if} \quad p^2 \mid n \text{ for some prime } p, \\ (-1)^k & \text{if} \quad m = p_1 p_2 \dots p_k \text{ for distinct primes } p_i. \end{cases}$$

For every $l, k \in \mathbb{N}$ it holds [21, p. 237] that

$$r_l(k) = \mu\left(\frac{l}{(l, k)} \right) \frac{\varphi(l)}{\varphi\left(\frac{l}{(l,k)} \right)}. \tag{3.18}$$

We can now explicitly evaluate the multinomial $\alpha_{\mathbb{Z}_m^*,d}(\boldsymbol{t})$, obtaining

$$\begin{aligned}\alpha_{\mathbb{Z}_m^*,d}(\boldsymbol{t}) &= \tfrac{1}{m}\sum_{1\le k\le m}\left(\sum_{l|m}\tfrac{1}{\varphi(l)}\sum_{j\in\mathbb{Z}_l^*}e^{\frac{2\pi}{l}jki}t_l\right)^d\\ &= \tfrac{1}{m}\sum_{1\le k\le m}\left(\sum_{l|m}\tfrac{1}{\varphi(l)}r_l(k)t_l\right)^d\\ &= \tfrac{1}{m}\sum_{k|m}\varphi\left(\tfrac{m}{k}\right)\left(\sum_{l|m}\frac{\mu\left(\frac{l}{(l,k)}\right)}{\varphi\left(\frac{l}{(l,k)}\right)}t_l\right)^d.\end{aligned}$$

It follows that the average type-spectrum of the (c,d)-regular $\mathbb{Z}_m^*$-labelled LDPC ensemble of $\mathbb{Z}_m$-codes is given by

(3.19)
$$\Gamma_{(\mathbb{Z}_m^*,c,d)}(\boldsymbol{\theta})=\mathrm{H}(\boldsymbol{\theta})+\frac{c}{d}\inf_{\boldsymbol{t}}\left\{\log\left(\frac{1}{m}\sum_{k|m}\varphi\left(\frac{m}{k}\right)\left(\sum_{l|m}\frac{\mu\left(\frac{l}{(l,k)}\right)}{\varphi\left(\frac{l}{(l,k)}\right)}t_l\right)^d\right)+dD\big(\pi_{\mathbb{Z}_m^*}\boldsymbol{\theta}||\boldsymbol{z}\big)\right\},$$

where the above infimum has to be considered with respect to all $\boldsymbol{t}$ in $\mathcal{P}(\mathbb{D}_m)$ such that $\mathrm{supp}(\boldsymbol{t})=\mathrm{supp}(\pi_{\mathbb{Z}_m^*}\boldsymbol{\theta})$. Of course, when m is prime, formula (3.19) reduces to (3.15). In particular, when $m=2$, (3.15), (3.17), and (3.19) coincide. For nonprime m instead, (3.19) is novel, to the best of our knowledge.

4. On low-weight type-spectra. In this section we will deal with estimations of the average type-spectra of the regular F-labelled LDPC G-code ensembles for G-types very close to the all-zero type δ_0. We will consider the variational distance on $\mathcal{P}(G)$, $||\boldsymbol{\theta}-\boldsymbol{\theta}'||:=\sup_{B\subseteq G}\{\boldsymbol{\theta}(B)-\boldsymbol{\theta}'(B)\}$.

Recall that, since we are dealing with LDPC G-codes, the all-zero N-tuple is always a codeword. Then $W_N(\delta_0)=1$ deterministically, i.e., for any realization of Π_N in the interconnection group $S_{Nc}\ltimes F^{Nc}$. Hence clearly $\Gamma_{(F,c,d)}(\delta_0)=0$. The main result of this section is that there exists a punctured neighborhood of δ_0 in $\mathcal{P}(G)$, over which the spectra $\Gamma_{(F,c,d)}(\boldsymbol{\theta})$ are strictly negative. Actually, much more precise results will be derived, characterizing the exact rate of decay (asymptotically in N) of the sum of the average enumerating coefficients over all G-types $\boldsymbol{\theta}$ such that $0<||\boldsymbol{\theta}-\delta_0||<\frac{2}{d}$.

Throughout this section we will often use the following notation: for a,t in $\mathbb{N}$ we define the discrete intervals $I_t^a:=[(t-1)a+1,ta]\cap\mathbb{N}$. Notice that, given a degree pair (c,d), for every admissible block-length N in $\mathcal{N}_{(c,d)}$ we have $\{1,2,\dots,Nc\}=\bigcup_{1\le t\le L}I_t^d=\bigcup_{1\le s\le N}I_s^c$.

4.1. An upper bound to low-weight spectra. We start by deriving an upper bound to low-weight type-enumerating coefficients for the inner encoder $|Z_{\boldsymbol{\theta}}^{i,N}|:=\left|G_{\boldsymbol{\theta}}^{Nc}\cap\ker\mathrm{Sum}_d^N\right|$.

LEMMA 4.1. *Let (c,d) be a degree pair, and let $N\in\mathcal{N}_{(c,d)}$. For every $\boldsymbol{\theta}$ in $\mathcal{P}_{Nc}(G)$ such that*

$$||\boldsymbol{\theta}-\delta_0||\le\frac{2}{d},\tag{4.1}$$

we have

$$\left|Z_{\boldsymbol{\theta}}^{i,N}\right|\le\binom{L}{\lfloor w/2\rfloor}\binom{\lfloor w/2\rfloor\, d}{w}\binom{w}{\boldsymbol{\omega}},\tag{4.2}$$

where $\boldsymbol{\omega} \in \mathbb{N}^{G\setminus\{0\}}$ *is defined by* $\boldsymbol{\omega}(k) := Nc\,\boldsymbol{\theta}(k)$*, and* $w := \sum_{k=1}^{m-1} \boldsymbol{\omega}(k)$ *is the number of nonzero entries in an* Nc*-tuple of type* $\boldsymbol{\theta}$*.*

Proof. Let $\boldsymbol{y}$ in $G_{\boldsymbol{\theta}}^{Nc}$ be any Nc-tuple of type $\boldsymbol{\theta}$. A necessary condition for $\boldsymbol{y}$ to be in $\ker \mathrm{Sum}_d^N$ is that each of the first L intervals I_t^d contains either none or at least two nonzero entries of $\boldsymbol{y}$. It follows from (4.2) that $\left|\left\{t \le L : \left|\mathrm{supp}(\boldsymbol{y}) \cap I_t^d\right| \ge 2\right\}\right| \le \lfloor w/2 \rfloor$, while, for any choice of a dissection $1 \le t_1 < \cdots < t_{\lfloor w/2 \rfloor} \le L$ (notice that (4.1) implies $w/2 \le L$), we have $|\{\boldsymbol{y} \in G_{\boldsymbol{\theta}}^{Nc} : \mathrm{supp}(\boldsymbol{y}) \subseteq \bigcup_{j=1}^{\lfloor w/2 \rfloor} I_{t_j}^d\}| \le \binom{d\lfloor w/2 \rfloor}{w}\binom{w}{\boldsymbol{\omega}}$. It follows that

$$\begin{aligned}
|Z_{\boldsymbol{\theta}}^{i,N}| &\le \left| \bigcup_{1\le t\le L} \left\{\boldsymbol{y} \in G_{\boldsymbol{\theta}}^{Nc} : \left|\mathrm{supp}(\boldsymbol{y}) \cap I_t^d\right| \ne 1\right\} \right| \\
&\le \left| \bigcup_{1\le t_1<\cdots<t_{\lfloor w/2 \rfloor}\le L} \left\{ \boldsymbol{y} \in G_{\boldsymbol{\theta}}^{Nc} : \mathrm{supp}(\boldsymbol{y}) \subseteq \bigcup_{j=1}^{\lfloor w/2 \rfloor} I_{t_j} \right\} \right| \\
&\le \binom{L}{\lfloor w/2 \rfloor}\binom{d\,\lfloor w/2 \rfloor}{w}\binom{w}{\boldsymbol{\omega}}. \qquad \square
\end{aligned}$$

We now obtain an estimation for the average low-weight type-enumerators.

LEMMA 4.2. *Let* (c,d) *be a degree pair,* $F \le \mathrm{Aut}(G)$*, and* $N \in \mathcal{N}_{(c,d)}$*. For every* $\boldsymbol{\theta} \in \mathcal{P}_N(G)$ *satisfying* (4.1) *the average type-enumerator function of the* (c,d)*-regular* F*-labelled ensemble satisfies*

$$\overline{W_N(\boldsymbol{\theta})} \le \binom{N}{N\boldsymbol{\theta}}\binom{L}{\lfloor w/2 \rfloor}\left(\frac{w}{2L}\right)^w, \tag{4.3}$$

where $w := Nc(1-\boldsymbol{\theta}(0))$*.*

Proof. Consider the projection map $\pi_F : G \to G/F$ and the associated map for types $\pi_F^\sharp : G \to G/F$. Define $\boldsymbol{\upsilon} := \pi_F^\sharp \boldsymbol{\theta}$, and $\boldsymbol{u} \in \mathbb{Z}_+^{G/F\setminus\{0\}}$ by $\boldsymbol{u}(k) = Nc\boldsymbol{\upsilon}(k)$. Also, for every $\boldsymbol{\theta}'$ in $\mathcal{P}(G)$, define $\boldsymbol{\omega}'$ in $\mathbb{Z}_+^{G\setminus\{0\}}$ by $\boldsymbol{\omega}'(k) := Nc\,\boldsymbol{\theta}'(k)$. Notice that $\sum_{\boldsymbol{\theta}' \in \mathcal{O}_{\boldsymbol{\upsilon}}^{Nc}} \binom{w}{\boldsymbol{\omega}'} = \binom{w}{Nc\,\boldsymbol{u}}\boldsymbol{\varphi}^{Nc\boldsymbol{\upsilon}}$. From (3.10), (3.13), and (4.2) we get

$$\begin{aligned}
\overline{W_N(\boldsymbol{\theta})} &= \binom{N}{N\boldsymbol{\theta}}\binom{Nc}{Nc\,\boldsymbol{\upsilon}}^{-1} \boldsymbol{\varphi}^{-Nc\,\boldsymbol{\upsilon}} \sum_{\boldsymbol{\theta}' \in \mathcal{O}_{\boldsymbol{\upsilon}}^{Nc}} |Z_{\boldsymbol{\theta}'}^{i,N}| \\
&\le \binom{N}{N\boldsymbol{\theta}}\binom{Nc}{w}^{-1}\binom{L}{\lfloor w/2 \rfloor}\binom{\lfloor w/2 \rfloor\, d}{w}\binom{w}{Nc\,\boldsymbol{u}}^{-1} \boldsymbol{\varphi}^{-Nc\,\boldsymbol{\upsilon}} \sum_{\boldsymbol{\theta}' \in \mathcal{O}_{\boldsymbol{\upsilon}}^{Nc}} \binom{w}{\boldsymbol{\omega}'} \\
&= \binom{N}{N\boldsymbol{\theta}}\binom{L}{\lfloor w/2 \rfloor}\binom{Nc}{w}^{-1}\binom{\lfloor w/2 \rfloor\, d}{w} \\
&= \binom{N}{N\boldsymbol{\theta}}\binom{L}{\lfloor w/2 \rfloor}\frac{\lfloor w/2 \rfloor d(\lfloor w/2 \rfloor d-1)\ldots(\lfloor w/2 \rfloor d-w+1)}{Ld(Ld-1)\ldots(Ld-w+1)} \\
&\le \binom{N}{N\boldsymbol{\theta}}\binom{L}{\lfloor w/2 \rfloor}\left(\frac{w}{2L}\right)^w. \qquad \square
\end{aligned}$$

A first consequence of Lemma 4.2 is the following upper bound on the average type-spectra of the F-labelled LDPC ensembles.

PROPOSITION 4.3. *For every degree pair* (c,d) *such that* $c \ge 3$ *we have*

$$\Gamma_{(F,c,d)}(\boldsymbol{\theta}) \le f_{c,d}(x) \qquad \forall\, \boldsymbol{\theta} : \|\boldsymbol{\theta} - \delta_0\| \le \frac{2}{d},$$

where $x := 1 - \boldsymbol{\theta}(0)$*, and*

$$f_{c,d}(x) := \mathrm{H}(x) + x\log(|G|-1) + \tfrac{c}{d}\,\mathrm{H}\left(\tfrac{d}{2}x\right) + cx\log\left(\tfrac{d}{2}x\right),$$

with $\mathrm{H}(x) := -x\log x - (1-x)\log(1-x)$ *denoting the binary entropy.*

Proof. From (4.3) it follows that, for every $||\boldsymbol{\theta} - \delta_0|| < \frac{2}{d}$, for the F-labelled (c,d)-regular ensemble we have

$$\begin{aligned}\frac{1}{N}\log\overline{W_N(\boldsymbol{\theta})} \quad &\leq \quad \frac{1}{N}\log\binom{N}{N\boldsymbol{\theta}} + \frac{1}{N}\log\binom{L}{\lfloor xN\frac{c}{2}\rfloor} + \frac{1}{N}\log\left(\frac{cNx}{2L}\right)^{cNx} \\ &\overset{N\in\mathcal{N}_{(c,d)}}{\longrightarrow} \ \mathrm{H}(\boldsymbol{\theta}) + \frac{c}{d}\,\mathrm{H}\left(\frac{d}{2}x\right) + +cx\log\left(\frac{d}{2}x\right) \\ &\leq \quad \mathrm{H}(x) + x\log(|G|-1) + cx\log\left(\frac{d}{2}x\right). \qquad \square\end{aligned}$$

It is easy to see that, whenever $c > 2$, $\frac{\mathrm{d}}{\mathrm{d}x} f_{c,d}\big|_{x=0} = -\infty$. Therefore, Proposition 4.3 implies that the spectra $\Gamma_{(F,c,d)}(\boldsymbol{\theta})$ are strictly negative in a sufficiently small punctured neighborhood of δ_0 in $\mathcal{P}(G)$. In section 5 this fact will be used in order to show that the minimum $\boldsymbol{\Delta}$-distance grows linearly with N with high probability. Here we derive more precise estimations for the average type-enumerating functions.

PROPOSITION 4.4. *Let F be any subgroup of* $\mathrm{Aut}(G)$*, (c,d) a degree pair, and $N \in \mathcal{N}_{(c,d)}$. There exists a positive constant K such that the type-enumerator function of the (c,d)-regular F-labelled ensemble satisfies*

$$\sum_{\frac{2}{N}\leq||\delta_0-\boldsymbol{\theta}||\leq\frac{2}{d}} \overline{W_N(\boldsymbol{\theta})} \leq KN^{2-c}.$$

Proof. For every N in $\mathcal{N}_{(c,d)}$ we define the quantities

$$g_w(N) := \sum_{||\delta_0-\boldsymbol{\theta}||=\frac{w}{N}} \overline{W_N(\boldsymbol{\theta})}, \qquad w \in \mathbb{N}.$$

For $\boldsymbol{\theta}$ in $\mathcal{P}_N(G)$ define $\boldsymbol{\omega}$ as in Lemma 4.1. For all $w = 2, \dots, \lfloor \frac{2}{d}N \rfloor$, (4.3) implies

$$g_w(N) \leq \sum_{\boldsymbol{\theta}(0)=\frac{N-w}{N}} \binom{N}{N\boldsymbol{\theta}}\binom{L}{\lfloor c\frac{w}{2}\rfloor}\left(\frac{wc}{2L}\right)^{wc} = \binom{L}{\lfloor c\frac{w}{2}\rfloor}\left(\frac{wc}{2L}\right)^{wc}\binom{N}{w}(|G|-1)^w =: g'_w(N).$$

We have, for every $2 \leq w \leq \lfloor 2dN \rfloor$,

$$\frac{g'_{w+2}(N)}{g'_w(N)} \leq (|G|-1)^2\left(\frac{N-w}{w}\right)^2\left(\frac{L-\lfloor c\frac{w}{2}\rfloor}{\lfloor c\frac{w}{2}\rfloor\, 2L}\right)^c\left(1+\frac{2}{w}\right)^{(w+2)c} \leq (|G|-1)^2(3e)^{2c}N^{2-c}.$$

It follows that if $c \geq 3$, then there exists N_0 in $\mathbb{N}$ such that, for all N in $\mathcal{N}_{(c,d)}$ such that $N \geq N_0$, $\frac{g'_{w+2}(N)}{g'_w(N)} \leq \frac{1}{2}$ for all $1 \leq w \leq \lfloor \frac{2}{d}N \rfloor$. Then we have

$$\sum_{\frac{2}{N}\leq||\delta_0-\boldsymbol{\theta}||\leq\frac{2}{d}} \overline{W_N(\boldsymbol{\theta})} \leq g'_2(N)\sum_{w=2}^{\lfloor\frac{2}{d}N\rfloor} 2^{-w} + g'_3(N)\sum_{w=2}^{\lfloor\frac{2}{d}N\rfloor} 2^{-w} \leq 2g'_2(N) + 2g'_3(N) \leq KN^{2-c}$$

for some positive constants K', K'', K, all independent of N. $\square$

4.2. On weight-one codewords. We now derive a more precise estimation of the average enumerating functions for G-types of N-tuples with all but one entry equal to zero. Fixed any N in $\mathbb{N}$, k in G we define the G-type

$$\boldsymbol{\tau}_k := \left(1 - \frac{1}{N}\right)\delta_0 + \frac{1}{N}\delta_k \in \mathcal{P}_N(G),$$

and we look for upper bounds on the average spectra $\overline{W_N(\boldsymbol{\tau}_k)}$ for the (c,d)-regular F-labelled LDPC ensembles. We will show how these estimations depend on the choice of F among the subgroups of the automorphism group $\mathrm{Aut}(G)$.

We start with a few elementary considerations about closed walks and cycles in directed graphs. A closed walk of length n in a directed graph $\mathcal{G} = (V, E)$ (where V is a finite set and $E \subseteq V^2$) is a $\mathbb{Z}_n$-labelled string of vertices $\boldsymbol{v} \in V^{\mathbb{Z}_n}$ such that any two consecutive vertices are adjacent, i.e., $(v_k, v_{k+1}) \in E$ for all $k \in \mathbb{Z}_n$. A cycle of length n is a closed walk $\boldsymbol{v} \in V^{\mathbb{Z}_n}$ such that $v_k \neq v_j$ for all $k \neq j \in \mathbb{Z}_n$. A self-loop is a cycle of length 1. Every closed walk $\boldsymbol{v}$ of length n is the concatenation of k cycles $\boldsymbol{v}^1, \dots, \boldsymbol{v}^k$ such that the sum of the lengths of $\boldsymbol{v}^1, \dots, \boldsymbol{v}^k$ equals n. Observe that in general $k \leq n$, while $k \leq \lfloor n/2 \rfloor$, provided that the directed graph $\mathcal{G}$ contains no self-loops.

Given a finite Abelian group G and a subset S of G, we denote by $\mathcal{G}(G, S)$ the directed Cayley graph with vertex set G and edge set $\{(g, g+s) \,|\, g \in G,\, s \in S\}$. It is straightforward that closed walks $\boldsymbol{v}$ of length n in an Abelian Cayley graph $\mathcal{G}(G, S)$ starting in any fixed vertex $g \in G$ (i.e., such that $v_0 = g$) are in one-to-one correspondence with 0-sum n-tuples $\boldsymbol{x}$ in S^n.

For a subset $S \subseteq G$ and a positive integer n, consider a closed walk $\boldsymbol{v}$ of length n in $\mathcal{G}$. By the previous considerations, $\boldsymbol{v}$ is the concatenation of $k(\boldsymbol{v})$ cycles. We put $b(S, n)$ equal to the maximum of $k(\boldsymbol{v})$ over all possible closed walks $\boldsymbol{v}$ of length n in $\mathcal{G}(G, S)$, with the agreement that $b(S, n) = 0$ whenever no closed walk in $\mathcal{G}(G, S)$ has length n. The reason for this notation becomes evident with the following result.

LEMMA 4.5. *Let F be any subgroup of* $\mathrm{Aut}(G)$, *(c, d) a degree pair, and $N \in \mathcal{N}_{(c,d)}$. Then, for all k in G, the type-enumerator function of the (c, d)-regular F-labelled ensemble satisfies*

$$\overline{W_N(\boldsymbol{\tau}_k)} \leq N \binom{L}{b(Fk, c)} \left[\frac{b(Fk, c)}{L}\right]^c. \tag{4.4}$$

Proof. Define $\boldsymbol{v} := \pi_F^\sharp \boldsymbol{\tau}_k \in \mathcal{P}(G/F)$. Let $\boldsymbol{y}$ be any element of $G_{\boldsymbol{v}}^{Nc}$. Then for $\mathrm{Sum}_d^N \boldsymbol{y} = \boldsymbol{0}$ in G^L it is necessary that $\sum_{1 \leq j \leq Nc} y_j = 0$ in G. Since $\boldsymbol{y} \in G_{\boldsymbol{v}}^{Nc}$ has exactly c nonzero entries all belonging to Fk, it follows that $\left|Z_{\boldsymbol{v}}^{i,N}\right| = 0$ iff there are no closed walks of length c in the Cayley graph $\mathcal{G}(G, Fk)$. Then (4.4) immediately follows in the case $b(Fk, c) = 0$.

Now assume that there exist closed walks of length c in $\mathcal{G}(G, Fk)$. By the previous considerations, each such walk decomposes in at most $b(Fk, c)$ cycles. If we consider the intervals I_t^d, for $1 \leq t \leq L$, and put $\mathrm{supp}(\boldsymbol{y}) \cap I_t^d := \{j_1^t, j_2^t, \dots, j_{n_t}^t\}$, we have

$$\left(\mathrm{Sum}_d^N \boldsymbol{y}\right)_t = \sum_{j \in I_t^d} y_j = \sum_{1 \leq i \leq n_t} y_{j_i^t} \qquad \forall\, 1 \leq t \leq L.$$

Therefore, if $\mathrm{Sum}_d^N \boldsymbol{y} = \boldsymbol{0}$, then it is necessary that $\boldsymbol{v} \in G^{\mathbb{Z}_{n_t}}$, $v_l := \sum_{1 \leq i \leq l} y_{j_i^t}$ is a closed walk in $\mathcal{G}(G, Fk)$ for all t such that $\mathrm{supp}(\boldsymbol{y}) \cap I_t^d$ is nonempty. It follows that

$\mathrm{supp}(\boldsymbol{y}) \cap I_t^d$ is nonempty for at most $b(Fk,c)$ values of t. Therefore, by taking into account the $\binom{L}{b(Fk,c)}$ possible choices of $b(Fk,c)$ intervals out of L possible ones, the $\binom{b(Fk,c)}{c}$ choices of c positions out of $b(Fk,c)d$ available ones, and the $\boldsymbol{\varphi}(Fk)^c$ choices of an ordered c-tuple with entries from the orbit Fk, we get

$$\left|Z_{\boldsymbol{v}}^{i,N}\right| = \left|\ker \mathrm{Sum}_d^N \cap G_{\boldsymbol{v}}^{Nc}\right| \leq \binom{L}{b(Fk,c)}\binom{b(Fk,c)d}{c}\boldsymbol{\varphi}(Fk)^c.$$

Then from (3.10) it follows that

$$\begin{aligned}\overline{W_N(\boldsymbol{\tau}_k)} = \frac{N\left|Z_{\boldsymbol{v}}^{i,N}\right|}{\binom{Nc}{c}\boldsymbol{\varphi}(Fk)^c} &\leq \frac{N}{\binom{Nc}{c}}\binom{L}{b(Fk,c)}\binom{b(Fk,c)d}{c} \\ &\leq N\binom{L}{b(Fk,c)}\left[\frac{b(Fk,c)}{L}\right]^c. \qquad \square\end{aligned}$$

4.3. Main result. Building on the results of sections 4.1 and 4.2, we are now ready to present the main result of this section. For a subgroup F of $\mathrm{Aut}(G)$ and a positive integer c we define

$$a(F,c) := 1 - c + \max\left(\{1\} \cup \{b(Fk,c) |\, k \in G \setminus \{0\}\}\right), \tag{4.5}$$

where we recall that $b(S,c)$ was defined in section 4.2 as the minimum number of cycles in $\mathcal{G}(G,S)$ of total length c, with the agreement that $b(S,c) = 0$ when no closed walk in $\mathcal{G}(G,S)$ has length c.

Before stating the main result, we need a simple property of $a(F,c)$. For every $k \neq 0$, Fk does not contain 0, so that there are no self-loops in $\mathcal{G}(G,Fk)$, and then $b(Fk,c) \leq \lfloor c/2 \rfloor$. It immediately follows that

$$2 - c \leq a(F,c) \leq 1 - \lceil c/2 \rceil. \tag{4.6}$$

THEOREM 4.6. *For every degree pair (c,d) such that $c \geq 3$, and every subgroup F of $\mathrm{Aut}(G)$, there exists a positive constant K such that for the (c,d)-regular F-labelled ensemble it holds that*

$$\sum_{0<||\delta_0 - \boldsymbol{\theta}||\leq \frac{2}{d}} \overline{W_N(\boldsymbol{\theta})} \leq K N^{a(F,c)}, \qquad N \in \mathcal{N}_{(c,d)}.$$

Proof. First, we consider weight-one types. From (4.4) we have

$$\sum_{\boldsymbol{\theta}(0)=\frac{N-1}{N}} \overline{W_N(\boldsymbol{\theta})} \leq \sum_{k\in G\setminus\{0\}} N\binom{L}{b(Fk,c)}\frac{b(Fk,c)^c}{L^c} \leq K' \sum_{k\in G\setminus\{0\}} N^{1+b(Fk,c)-c} \leq K'|G|N^{a(F,c)}$$

for some positive constant K'. The claim then follows by combining Proposition 4.4 with the previous estimation and observing that $a(F,c) \leq 2 - c \leq -1$. $\square$

Now we explicitly evaluate $a(F,c)$ for the three examples studied in the previous section.

Example 4. Consider the case when $G \simeq \mathbb{Z}_p^r$ and either $F \simeq \mathrm{Aut}(\mathbb{Z}_p^r)$ or $F \simeq \mathbb{F}_{p^r}^*$. In both cases $Fk = \mathbb{Z}_p^r \setminus \{0\}$ for all $k \in \mathbb{Z}_p^r \setminus \{0\}$. Then $\mathcal{G}(\mathbb{Z}_p^r, Fk) = \mathcal{G}(\mathbb{Z}_p^r, \mathbb{Z}_p^r \setminus \{0\})$ is the complete graph with p^r vertices. It follows that $\mathcal{G}(\mathbb{Z}_p^r, \mathbb{Z}_p^r \setminus \{0\})$ contains closed walks of any length $n \geq 2$ whenever $p^r \neq 2$, while $\mathcal{G}(\mathbb{Z}_2, \{1\})$ contains closed walks

of even length only. Therefore, for $G \simeq \mathbb{Z}_p^r$ with $p^r \neq 2$, $a(F,c) = 1 - \lceil c/2 \rceil$ for all c, while for $G \simeq \mathbb{Z}_2$, $a(F,c) = 1 - c/2$ for even c and $2 - c$ for odd c.

Example 5. Consider the unlabelled ensemble over the cyclic group, i.e., $G \simeq \mathbb{Z}_m$ with $F = \{1\}$. If $(m,c) = 1$, then $m|ck$ iff $m|k$. Then, for all $k \in \mathbb{Z}_m \setminus \{0\}$, the Cayley graph $\mathcal{G}(\mathbb{Z}_m, Fk) = \mathcal{G}(\mathbb{Z}_m, \{k\})$ has no closed walks of length c. In this case clearly $a(\{1\}, c) = 2 - c$.

Then consider the case when $(m,c) > 1$, and let $\mathrm{lpcf}(c,m)$ be the smallest prime common factor between c and m. Consider any k in $\mathbb{Z}_m \setminus \{0\}$ such that $\mathcal{G}(\mathbb{Z}_m, \{k\})$ has a closed walk of length c, i.e., such that $m \mid ck$. The length of the shortest such walk is given by $\frac{m}{(m,k)} = \frac{(m,ck)}{(m,k)} = (\frac{m}{(m,k)}, c)$. Thus, $\frac{m}{(m,k)} \mid c$, while clearly $\frac{m}{(m,k)} \mid m$. But $(m,k) < m$, so that necessarily the shortest cycle in $\mathcal{G}(\mathbb{Z}_m, \{k\})$ $\frac{m}{(m,k)}$ is not less than $\mathrm{lpcf}(m,c)$, with equality iff $k \in \frac{m}{\mathrm{lpcf}(m,c)}\mathbb{Z}_m \setminus \{0\}$. Thus, $b(\{k\}, c) = \frac{c}{\mathrm{lpcf}(m,c)}$ for $k \in \frac{m}{\mathrm{lpcf}(m,c)}\mathbb{Z}_m \setminus \{0\}$, and $b(\{k\}, c) < \frac{c}{\mathrm{lpcf}(m,c)}$ for $k \in \mathbb{Z}_m \setminus \frac{m}{\mathrm{lpcf}(m,c)}\mathbb{Z}_m$. It immediately follows that, whenever $(m,c) > 1$, $a(\{1\}, c) = 1 - c + \frac{c}{\mathrm{lpcf}(m,c)}$.

Example 6. Consider the uniformly labelled ensemble over the cyclic group, i.e., $G \simeq \mathbb{Z}_m$ with $F \simeq \mathbb{Z}_m^*$. First, we claim, for $n \geq 2$, the following:

- if n is even, then all closed walks in $\mathcal{G}(\mathbb{Z}_n, \mathbb{Z}_n^*)$ have even length and there exists a 2-cycle;
- if n is odd, then there exist both a 2-cycle and a 3-cycle.

To see this, first, since $1, -1 \in \mathbb{Z}_n^*$, $(0,1)$ is a 2-cycle in $\mathcal{G}(\mathbb{Z}_n, \mathbb{Z}_n^*)$, both for even and odd n. Then consider the case when n is even: clearly all $k \in \mathbb{Z}_n^*$ are odd, so that the modulo-n sum of an odd number of elements of $\mathbb{Z}_n^*$ cannot be equal to 0 modulo n. Thus every closed walk in $\mathcal{G}(\mathbb{Z}_n, \mathbb{Z}_n^*)$ must be of even length. On the other hand, if n is odd, then $2 \in \mathbb{Z}_n^*$, so that $(0,2,1)$ is a 3-cycle in $\mathcal{G}(\mathbb{Z}_n, \mathbb{Z}_n^*)$.

Let us now consider some $k \in \mathbb{Z}_m \setminus \{0\}$. Then, by applying the previous observation with $n = \frac{m}{(m,k)}$, one gets that, if c is odd and $\frac{m}{(m,k)}$ is even, there are no closed walks of length c in $\mathcal{G}(\mathbb{Z}_m, \mathbb{Z}_m^* k)$ so that $b(\mathbb{Z}_m^* k, c) = 0$, while otherwise, if c is even or $\frac{m}{(m,k)}$ is odd, $b(\mathbb{Z}_m^* k, c) = \lfloor c/2 \rfloor$. It thus follows that $a(\mathbb{Z}_m^*, c) = 1 - \lceil c/2 \rceil$ unless c is odd and m is an integer power of 2; in the latter case $a(\mathbb{Z}_m^*, c) = 2 - c$.

4.4. Lower bounds on low-weight type-enumerators. In this section we present some results, of independent interest, which show that the estimations given by Theorem 4.6 are tight. All of the proofs are deferred to the appendix.

First, we deal with weight-one type-enumerators.

PROPOSITION 4.7. *Let (c,d) be a degree pair such that $c \geq 3$, and let F be any subgroup of* $\mathrm{Aut}(G)$. *Then there exists a constant $K > 0$ such that for all k in $G \setminus \{0\}$ such that $a(F,c) = 1 - c + b(Fk, c)$ the type-enumerator function of the (c,d)-regular F-labelled LDPC ensemble satisfies*

$$\mathbb{P}\left(W_N(\boldsymbol{\tau}_k) \geq 1\right) \geq K N^{a(F,c)}, \qquad N \in \mathcal{N}_{(c,d)}. \tag{4.7}$$

Finally, we propose a lower bound on weight-two type-enumerators. For every k in G define

$$\hat{\boldsymbol{\tau}}_k := \frac{1}{N}\delta_k + \frac{1}{N}\delta_{-k} + \frac{N-2}{N}\delta_0 \in \mathcal{P}(G).$$

PROPOSITION 4.8. *For every degree pair (c,d) there exists a constant $K > 0$ such that for every k in $G \setminus \{0\}$ the type-enumerator function of the (c,d)-regular F-labelled LDPC ensemble satisfies*

$$\mathbb{P}\left(W_N(\hat{\boldsymbol{\tau}}_k) \geq 1\right) \geq K N^{2-c} \qquad \forall\, N \in \mathcal{N}_{(c,d)}. \tag{4.8}$$

5. Asymptotic lower bounds on the typical minimum distance. Throughout this section we will assume we have fixed a G-symmetric MC $(\mathcal{X}, \mathcal{Y}, P)$ with associated Bhattacharyya distance $\boldsymbol{\Delta}$ and weight $\boldsymbol{\delta}$, and we study the asymptotics of the minimum $\boldsymbol{\Delta}$-distance of regular LDPC G-code ensembles.

Given a degree pair (c, d), a natural candidate for the typical normalized minimum $\boldsymbol{\Delta}$-distance of the (c, d)-regular F-labelled ensemble is the quantity

$$\gamma_{(F,c,d)} := \inf\left\{\langle \boldsymbol{\delta}, \boldsymbol{\theta}\rangle \,\middle|\, \boldsymbol{\theta} \in \mathcal{P}(G) \setminus \{\delta_0\} \text{ s.t. } \Gamma_{(F,c,d)}(\boldsymbol{\theta}) \geq 0\right\}. \tag{5.1}$$

It turns out that $\gamma_{(F,c,d)}$ actually is a lower bound on the asymptotic normalized minimum distance for the (c, d)-regular F-labelled ensemble. This does not follow directly from Theorem 3.5 since $\lim_{\boldsymbol{\theta}\to\delta_0} \Gamma_{(F,c,d)}(\boldsymbol{\theta}) = 0$. However, using both Theorems 3.5 and 4.6 the following result can be proved.

THEOREM 5.1. *Let (c, d) be a degree pair such that $a(F, c) < -1$. Then for the (c, d)-regular F-labelled LDPC ensemble the following holds:*

$$\mathbb{P}\left(\liminf_{N\in\mathcal{N}_{(c,d)}} \tfrac{1}{N} \mathrm{d}_{\min}(\ker \Phi_N) \geq \gamma_{(F,c,d)}\right) = 1.$$

Proof. By (2.3) we have that

$$\frac{1}{N}\mathrm{d}_{\min}(\ker \Phi_N) = \inf\left\{\langle \boldsymbol{\delta}, \boldsymbol{\theta}\rangle \,\middle|\, \boldsymbol{\theta} \in \mathcal{P}(G) \setminus \{\delta_0\} \text{ s.t. } W_N(\boldsymbol{\theta}) \geq 1\right\} = \min\left\{\kappa'_N, \kappa''_N\right\},$$

where for every N in $\mathcal{N}_{(c,d)}$ we define

$$\begin{aligned}\kappa'_N &:= \inf\left\{\langle \boldsymbol{\delta}, \boldsymbol{\theta}\rangle \,\middle|\, 0 < ||\boldsymbol{\theta} - \delta_0|| < \tfrac{2}{d} \,:\, W_N(\boldsymbol{\theta}) \geq 1\right\},\\ \kappa''_N &:= \inf\left\{\langle \boldsymbol{\delta}, \boldsymbol{\theta}\rangle \,\middle|\, ||\boldsymbol{\theta} - \delta_0|| \geq \tfrac{2}{d} \,:\, W_N(\boldsymbol{\theta}) \geq 1\right\}.\end{aligned}$$

Clearly, $\liminf_N \frac{1}{N}\mathrm{d}_{\min}(\ker \Phi_N) = \min\{\rho', \rho''\}$, where we put $\rho' := \liminf_N \kappa'_N$ and $\rho'' := \liminf_N \kappa''_N$.

We start by establishing a lower bound on ρ''. Define $\Omega := \left\{\boldsymbol{\theta} : ||\boldsymbol{\theta} - \delta_0|| \geq \frac{2}{d}\right\}$ and, for each x in $\mathbb{R}$, the set

$$\Omega_x := \left\{\boldsymbol{\theta} \in \Omega \cap \mathcal{P}_{\mathbb{N}}(G) \text{ s.t. } \Gamma_{(F,c,d)}(\boldsymbol{\theta}) < x\right\}. \tag{5.2}$$

Now consider the quantity $\eta(x) := \inf\left\{\langle \boldsymbol{\delta}, \boldsymbol{\theta}\rangle \,\middle|\, \boldsymbol{\theta} \in \Omega \setminus \Omega_x\right\}$. Since $\Gamma_{(F,c,d)}(\boldsymbol{\theta})$ is an upper semicontinuous function of $\boldsymbol{\theta}$ and Ω is a closed subset of $\mathcal{P}(G)$, standard analytical arguments (see Lemma 8.1 in the appendix) allow us to conclude that η is a nondecreasing and lower semicontinuous function.

Let us now fix some arbitrary $\varepsilon > 0$. By successively applying a union bound estimation, the Markov inequality, Theorem 3.5, and (5.2), we get

$$\mathbb{P}\left(\bigcup_{\boldsymbol{\theta}\in\Omega_{-\varepsilon}} \{W_N(\boldsymbol{\theta}) \geq 1\}\right) \leq \sum_{\boldsymbol{\theta}\in\Omega_{-\varepsilon}} \mathbb{P}\left(W_N(\boldsymbol{\theta}) \geq 1\right) \leq \sum_{\boldsymbol{\theta}\in\Omega_{-\varepsilon}} \overline{W_N(\boldsymbol{\theta})} \leq \exp(-N(\varepsilon - f(N))),$$

with $\lim_N f(N) = 0$. It follows that $\sum_N \mathbb{P}\big(\bigcup_{\boldsymbol{\theta}\in\Omega_{-\varepsilon}} \{W_N(\boldsymbol{\theta}) \geq 1\}\big) < +\infty$, and thus the Borel–Cantelli lemma implies that with probability one the event $\bigcup_{\boldsymbol{\theta}\in\Omega_{-\varepsilon}} \{W_N(\boldsymbol{\theta}) \geq 1\}$ occurs only for finitely many N in $\mathcal{N}_{(c,d)}$. Hence,

$$\mathbb{P}\left(\rho'' < \eta(-\varepsilon)\right) \leq \mathbb{P}\left(\left\{\bigcup_{\boldsymbol{\theta}\in\Omega_{-\varepsilon}} \{W_N(\boldsymbol{\theta}) > 0\}\right\} \text{ i. o. } N \in \mathcal{N}_{(c,d)}\right) = 0 \qquad \forall\, \varepsilon > 0,$$

where i. o. stands for infinitely often. Notice that $\gamma_{(F,c,d)} = \eta(0)$. Hence, monotonicity and lower semicontinuity of the function η allow us to conclude that

(5.3)
$$\mathbb{P}\left(\rho'' < \gamma_{(F,c,d)}\right) = \mathbb{P}\left(\rho'' < \eta(0)\right) \leq \mathbb{P}\left(\rho'' < \lim_k \eta\left(-\tfrac{1}{k}\right)\right) = \lim_k \mathbb{P}\left(\rho'' < \eta\left(-\tfrac{1}{k}\right)\right) = 0.$$

Now let us consider the term ρ'. By sequentially applying a union bound estimation, the Markov inequality, and Theorem 4.6, we get for every N in $\mathcal{N}_{(c,d)}$

$$\mathbb{P}\left(\bigcup_{0<||\boldsymbol{\theta}-\delta_0||<\frac{2}{d}} \{W_N(\boldsymbol{\theta}) \geq 1\}\right) \leq \sum_{0<||\boldsymbol{\theta}-\delta_0||<\frac{2}{d}} \overline{W_N(\boldsymbol{\theta})} \leq K N^{a(F,c)}, \tag{5.4}$$

where K is a positive constant independent of N. Since $a(F,c) < -1$, we get

$$\sum_N \mathbb{P}\left(\bigcup_{0<||\boldsymbol{\theta}-\delta_0||<\frac{2}{d}} \{W_N(\boldsymbol{\theta}) \geq 1\}\right) \leq K \sum_N N^{a(F,c)} < +\infty.$$

By the Borel–Cantelli lemma we get that the event $\bigcup_{0<||\boldsymbol{\theta}-\delta_0||<\frac{2}{d}} \{W_N(\boldsymbol{\theta}) \geq 1\}$ occurs only for finitely many N in $\mathcal{N}_{(c,d)}$ with probability one. This yields $\mathbb{P}(\rho' = +\infty) = 1$, which, together with (5.3), implies the claim. ∎

We have proved the previous theorem under the assumption $a(F,c) < -1$. In fact, for $c = 2$ it is known, since Gallager's work [19], that deterministically the minimum distance cannot grow faster than logarithmically with the block-length N. From (4.6) it follows that if $c \geq 5$, then $a(F,c) < -1$ for any F, and if $c = 3$, then $a(F,c) = -1$ for any F, while, when $c = 4$, $a(F,c) < -1$ for some choices of F. However, one can weaken the assumption $a(F,c) < -1$ requiring only that $a(F,c) < 0$ (thus including the cases $c = 3$ and $c = 4$ for some F). In these cases, $\gamma_{(F,c,d)}$ still gives an asymptotic lower bound for the normalized minimum distances $\frac{1}{N}\,\mathrm{d}_{\min}(\ker \Phi_N)$ in a weaker probabilistic sense. In fact, a more detailed analysis enlightens a nonconcentration phenomenon. In order to describe it, first, for every degree pair (c,d) and every subgroup F of $\mathrm{Aut}(G)$, we define the following quantity:

(5.5)
$$\zeta_{(F,c)} := \begin{cases} \min\{\boldsymbol{\delta}(k) |\, k \in G \setminus \{0\} : a(F,c) = 1 - c + b(Fk,c)\} & \text{if} \quad a(F,c) \neq 2-c, \\ \min\{(2 - b(Fk,c))\boldsymbol{\delta}(k) |\, k \in G \setminus \{0\}\} & \text{if} \quad a(F,c) = 2-c. \end{cases}$$

We have the following result.

THEOREM 5.2. *Let (c,d) be a degree pair such that $a(F,c) = -1$. Then*

$$\lim_{N \in \mathcal{N}_{(c,d)}} \mathbb{P}\left(\frac{1}{N}\,\mathrm{d}_{\min}(\ker \Phi_N) \geq \gamma_{(F,c,d)}\right) = 1.$$

Moreover, if the random variables Π_N defining the (c,d)-regular unlabelled LDPC ensemble are mutually independent, we have

$$\mathbb{P}\left(\liminf_{N \in \mathcal{N}_{(c,d)}} \mathrm{d}_{\min}(\ker \Phi_N) = \zeta_{(F,c)}\right) = 1.$$

Theorem 5.2 is proved in the appendix. The probabilistic interpretation is as follows. In the case $a(F,c) = -1$, with probability one, the sequence of the unnormalized minimum distances $(\mathrm{d}_{\min}(\ker \Phi_N))$ contains a subsequence converging to $\zeta_{(F,c)}$.

Thus, while with increasing probability the minimum $\mathbf{\Delta}$-distance is growing linearly with the block-length N, almost surely a subsequence with constant minimum distance shows up. We observe that, for irregular binary LDPC ensembles, even more evident nonconcentration phenomena are known to arise; see [15, 31].

6. Numerical results. In this section we present some numerical results for the minimum distances of the LDPC ensembles which have been studied in this paper. We focus on a particular channel, the $\mathbb{Z}_8$-symmetric 8-PSK AWGN channel, and we compare the average distance-spectra of the regular unlabelled and uniformly labelled LDPC $\mathbb{Z}_8$-code ensembles. Our results indicate a strong superiority of the uniformly labelled (i.e., the one with label group $F \simeq \mathbb{Z}_8^*$) ensemble with respect to the unlabelled one (i.e., $F = \{1\}$). Then we compare these results with some contradicting analysis of the average error probability of these ensembles and discuss how this seeming paradox can be explained by invoking so-called expurgation arguments.

6.1. Numerical results for the average distance-spectra. Let us start with some general considerations. Suppose we are given any ensemble of G-codes with average type-spectrum $\Gamma(\boldsymbol{\theta})$. Let $\gamma := \inf\{\langle \boldsymbol{\theta}, \boldsymbol{\delta}\rangle | \, \boldsymbol{\theta} \in \mathcal{P}(G) \setminus \{\delta_0\} \text{ s.t. } \Gamma(\boldsymbol{\theta}) \geq 0\}$ be its designated typical normalized minimum distance which we are interested in computing. Notice that Γ is a map defined over the $(|G|-1)$-dimensional simplex $P(G)$ and therefore in general of difficult visualization whenever $|G| > 2$. It is then convenient and natural to introduce the average distance-spectrum as a one-dimensional projection of Γ:

(6.1)
$$\Upsilon : [0, \max\{\boldsymbol{\delta}(x) | \, x \in G\}] \to [-\infty, +\infty), \qquad \Upsilon(t) := \sup\left\{\Gamma(\boldsymbol{\theta}) \middle| \, \boldsymbol{\theta} \in \mathcal{P}(G) : \langle \boldsymbol{\delta}, \boldsymbol{\theta}\rangle = t\right\}.$$

It is immediate to verify that $\gamma = \inf\{t \in [0, \max\{\boldsymbol{\delta}(x) | \, x \in G\}] \, : \Upsilon(t) \geq 0\}$. Notice also that, for $|G| = 2$ and $|G| = 3$, all Bhattacharyya distances are proportional to the Hamming distance, so that the average distance spectrum Υ is independent (up to a rescaling factor) of the chosen G-symmetric channel. For $|G| \geq 4$ instead, Υ really depends on the choice of the Bhattacharyya distance $\mathbf{\Delta}$.

In Figure 6.1 the average distance-spectra of two regular LDPC $\mathbb{Z}_8$-code ensembles are reported. We considered the Bhattacharyya distance $\mathbf{\Delta}$ of the 8-PSK AWGN channel and normalized it in such a way that $\max\{\boldsymbol{\delta}(x) | \, x \in \mathbb{Z}_8\} = \mathbf{\Delta}(0, 4) = 1$. In each picture a degree pair (c, d) is fixed. The dash-dotted curve is the graph of the distance-spectrum $\Upsilon_{(\{1\},c,d)}(t)$ of the (c, d)-regular unlabelled LDPC ensemble, while the solid curve is the graph of the distance-spectrum $\Upsilon_{(\mathbb{Z}_8^*,c,d)}(t)$ of the (c, d)-regular uniformly labelled LDPC ensemble.

As a reference two dotted curves are also plotted in each picture. The one taking the value 0 for $t = 0$ is the distance spectrum of the binary (c, d)-regular LDPC ensemble $\Upsilon^2_{(c,d)}(t)$. It is straightforward to check that it is a lower bound for the distance spectrum of any $\mathbb{Z}_8$-LDPC ensemble: it suffices to restrict the optimization in (6.1) to $\mathbb{Z}_8$-types $\boldsymbol{\theta}$ supported on the binary subgroup $4\mathbb{Z}_8$.

The second dotted curve instead, taking value $\frac{1}{2}\log\frac{1}{2}$ for $t = 0$, corresponds to the distance-spectra of the $\mathbb{Z}_8$-code ensemble (with no sparsity constraints) of the same rate $R = \frac{1}{2}\log 8$. This ensemble is defined as a sequence of kernels of random homomorphisms $(\ker \Phi_N)$, each Φ_N being uniformly distributed over $\mathrm{Hom}(\mathbb{Z}_8^N, \mathbb{Z}_8^{N/2})$, the group of all homomorphisms from $\mathbb{Z}_8^N$ to $\mathbb{Z}_8^{N/2}$, with no sparsity constraint. $\mathbb{Z}_8$-code ensembles of codes are a natural generalization of the traditional linear-coding ensembles over finite fields [20, 2] and have been considered in [10] and [11] in order to characterize the capacity achievable by Abelian group codes over symmetric channels.

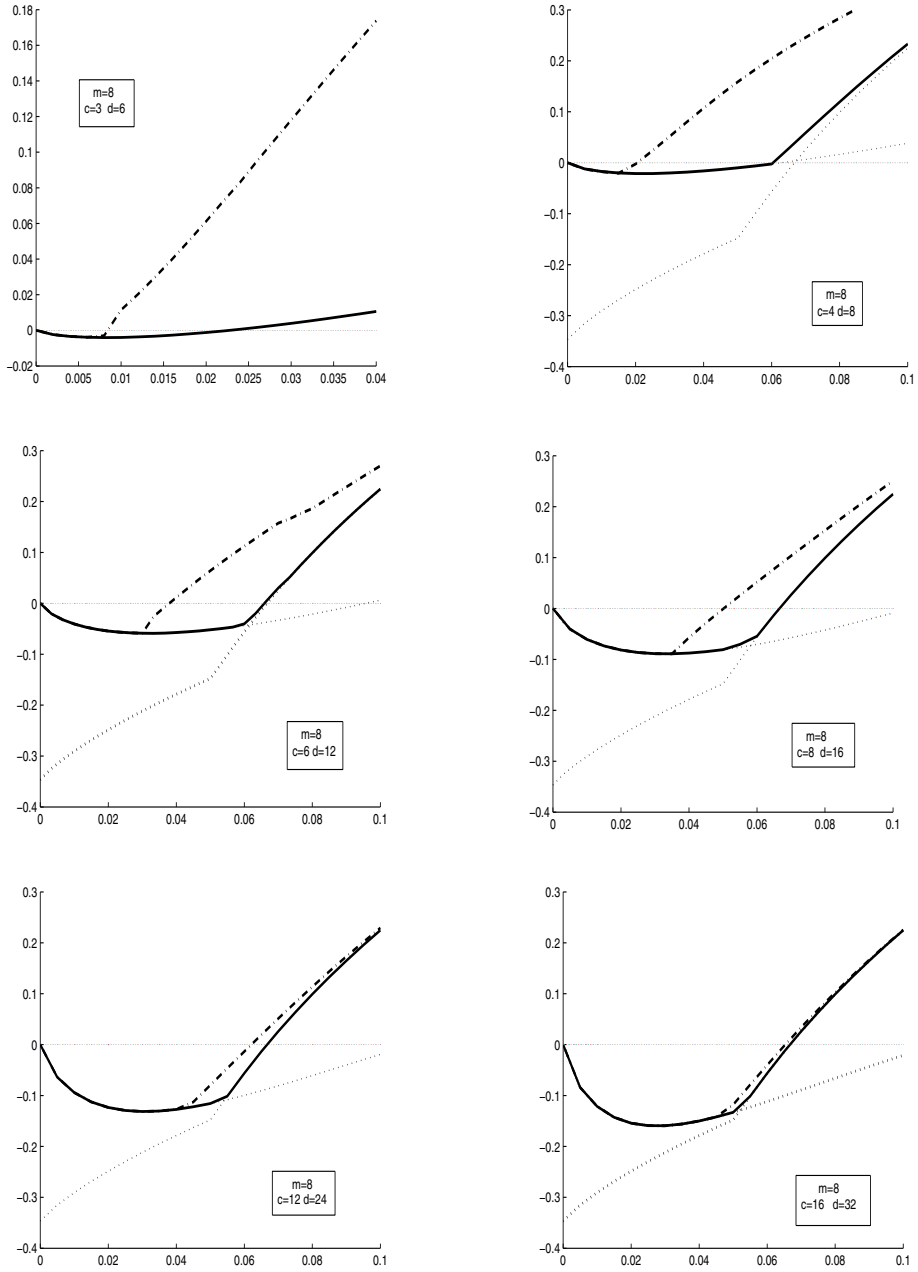


FIG. 6.1. *Bhattacharyya distance spectra of* (c, d)*-regular LDPC ensembles over* $\mathbb{Z}_8$ *for the* 8*-PSK AWGN channel: the solid curve corresponds to the uniformly labelled ensemble, the dash-dotted one corresponds to the unlabelled ensemble, and the two dotted curves correspond, respectively, to the* $\mathbb{Z}_8$*-linear ensemble and to the binary LDPC ensemble.*

In [12] their average type-spectra have been characterized; for the $\mathbb{Z}_8$-code ensemble of rate $\frac{1}{2}\log 8$ this is given by

$$\Gamma_{\mathbb{Z}_8}(\boldsymbol{\theta}) := \mathrm{H}(\boldsymbol{\theta}) - \frac{1}{2}\log l_8(\boldsymbol{\theta}), \qquad l_8(\boldsymbol{\theta}) := \frac{8}{\gcd\left(\mathrm{supp}\left(\boldsymbol{\theta}\right)\right)}.$$

Notice that $\Gamma_{\mathbb{Z}_8}(\boldsymbol{\theta})$ is an upper semicontinuous function over the simplex $\mathcal{P}(\mathbb{Z}_8)$, and its discontinuities correspond to types supported on the subgroups $2\mathbb{Z}_8$ and $4\mathbb{Z}_8$. In fact a salient point is easily recognizable in the graphs reported around the abscissa $t = 0.05$, corresponding to the intersection between the average spectrum of the binary subchannel and that of the $\mathbb{Z}_8$-subchannel. This salient point occurs before the curve crosses the t-axis, which is coherent with the fact, proved in [12], that the typical normalized minimum distance of the $\mathbb{Z}_8$-code ensemble equals the corresponding Gilbert–Varshamov bound. In other words, while for low values of t the distance spectrum of the $\mathbb{Z}_8$-code ensemble is dominated by the term corresponding to the smallest nontrivial subgroup (a phenomenon generally observable for Abelian group code ensembles), the value of the typical minimum distance is determined by types which are not supported in any proper subgroup of $\mathbb{Z}_8$ (this is instead related to the particular constellation chosen, although it is conjectured to be true for many constellations of interest).

Analogous considerations can be made about the LDPC distance-spectra based on the simulations reported. In particular, for distances close to 0, the average distance-spectra of both the unlabelled and the uniformly labelled $\mathbb{Z}_8$-LDPC ensembles are dominated by the binary-subgroup supported types. However, these components do affect the value of the typical normalized minimum distances ($\gamma_{(\{1\},c,d)}$ and $\gamma_{(\mathbb{Z}_8^*,c,d)}$, respectively) only for low values of the degrees ($c = 3, 4$). For all of the other values of the parameters, the typical minimum distance is instead determined by types which are not supported in any proper subgroup of $\mathbb{Z}_8$. Another observation which can be made is that, not surprisingly, as the values of the degrees (c, d) are increased while keeping their ratio constant, the distance-spectra of both the unlabelled and the uniformly labelled ensembles approach the one of the $\mathbb{Z}_8$-linear ensemble.

However, the most important conclusion which can be drawn from the graphics reported concerns the different behaviors of the unlabelled and the uniformly labelled ensembles. Indeed, it appears evident that the latter drastically outperforms the former at the distance level. In particular, already for relatively low values of the degrees ($c = 8$, $d = 16$) the uniformly labelled ensemble typical minimum distance $\gamma_{(\mathbb{Z}_8^*,c,d)}$ is very close (practically equal) to the Gilbert–Varshamov bound. For the same values of the degrees instead, the unlabelled ensemble suffers from a remarkable gap; this gap seems to be slowly filled up as the values of the degrees are increased, but it still remains significant for relatively high values of c and d. This indicates that structural properties of these two ensembles are remarkably different. Some prudence is nevertheless justified by the fact that ours are only lower bounds on the typical asymptotic normalized minimum distance, while, as already mentioned in the introduction, a concentration result for the type-spectra is needed in order to prove their tightness. However, while this phenomenon appears here only at the distance level, computer simulations of the performance of these codes reveal that a drastic superiority of the labelled ensemble with respect to the unlabelled one is evident also under belief-propagation decoding. We observe that this is coherent with Monte Carlo simulations reported in [4], where the labelled ensemble was shown to be closer to capacity than the unlabelled ensemble.

6.2. The average word error probability of the LDPC codes ensembles. In our analysis of the minimum distance properties of LDPC G-code ensembles, the quantities $\zeta_{(F,c)}$ show up as an almost sure lim inf for the unnormalized minimum distance only when $a(F,c) = -1$. However, these quantities characterize the asymptotic ML average performance of these ensembles for all values of $a(F,c)$.

For instance, let us consider in some detail the case $G \simeq \mathbb{Z}_{p^r}$ for some prime p and some positive integer r. Let us fix an admissible degree pair (c,d), and denote by $\overline{p_e(\mathcal{C}_N)}^{(F,c,d)}$ the average ML error probability of the (c,d)-regular F-labelled ensemble of LDPC $\mathbb{Z}_{p^r}$-codes over an arbitrary $\mathbb{Z}_{p^r}$-symmetric MC. Then it is possible to show that there exists a threshold $(1-\frac{c}{d})\log p^r < C_{(F,c,d)} < \log p^r$ such that, for every $\mathbb{Z}_{p^r}$-symmetric channel whose $\mathbb{Z}_{p^r}$-capacity (2.5) exceeds $C_{(F,c,d)}$, the average error probability $\overline{p_e(\mathcal{C}_N)}^{(F,c,d)}$ goes to zero in the limits of large N. Moreover, if one considers an increasing sequence of degree pairs (c_n,d_n) with a given designed rate $(1-\frac{c_n}{d_n})\log p^r$ converging to R, then the corresponding LDPC thresholds $C_{(c_n,d_n,F)}$ converge to R.

More precisely, it is possible to show that over any $\mathbb{Z}_{p^r}$-symmetric channel whose $\mathbb{Z}_{p^r}$-capacity exceeds $C_{(F,c,d)}$ we have

$$K_1 N^{a(F,c)} \leq \overline{p_e(\mathcal{C}_N)}^{(F,c,d)} \leq K_2 N^{a(F,c)} \tag{6.2}$$

for some positive constants K_1, K_2 both independent of N. Moreover, it can be proved that

$$\limsup_{N\in\mathcal{N}_{(c,d)}} \frac{\overline{p_e(\mathcal{C}_N)}^{(F,c,d)}}{N^{a(F,c)}} \leq K_3 \exp(\zeta_{(F,c)}) \tag{6.3}$$

for some positive constants K_3 independent of the channel (and thus from $\mathbf{\Delta}$). The results (6.2) are known in the binary case (see [29]); (6.2) was presented in [10] for the unlabelled LDPC ensemble. Proofs of (6.2), (6.3) in their full generality can be gathered coupling the estimations of section 4 with the standard bounding techniques used in [28, 39, 29, 4] and will be given elsewhere.

Observe that if $F \leq F' \leq \mathrm{Aut}(G)$, then

$$a(F,c) \leq a(F',c), \qquad \zeta_{(F,c)} \geq \zeta_{(F',c)}. \tag{6.4}$$

Thus, from the point of view of the average performance, the smaller the label group, the better the parameters. This stands in contrast with the numerical results presented in the previous paragraph, indicating that at the distance level the uniformly labelled ensembles perform much better than their unlabelled counterparts. An explanation for this seeming paradox can be obtained by invoking so-called expurgation arguments. Indeed, it can be proved that, while the average error probability of the LDPC ensembles is affected by a vanishingly small fraction of codes with low minimum distance and decays to zero only as a negative power of N, almost surely a sequence of codes sampled from the same ensemble has error probability decreasing to zero exponentially fast with N. It is this typical exponential behavior that has to be considered representative of the ensemble, rather than the one of the average error probability. It is also worth mentioning that the typical error exponent can be estimated in terms of the average type-spectra, using techniques presented in [39]. This phenomenon is well known in the LDPC code literature [19, 29]; proofs for LDPC codes over Galois fields can be found in [17, 4].

7. Conclusions. The following issues are left for future research:

- proving concentration results for the spectra of the LDPC ensembles for instance using a second-order method (see [33]);
- giving an analytical explanation of the different behavior of the labelled and unlabelled ensembles;
- generalizing the analysis to irregular ensembles following the approach of [15, 31];
- considering generalizations of the so-called stopping sets and pseudoweight distributions which in the binary case characterize the iterative decoding performance of LDPC codes (see [31, 43, 24]); while the distribution of stopping sets has been studied for binary LDPC ensembles, the distribution of pseudocodewords is unknown even in the binary case.

8. Appendix.

8.1. A semicontinuity lemma. Let E be a compact metric space. It is a standard fact that any lower semicontinuous function $f : E \to (-\infty, +\infty]$ achieves its minimum on every closed nonempty subset C of E, i.e.,

$$\exists\ \overline{x} \in C \text{ s.t. } f(\overline{x}) \leq f(x) \qquad \forall\ x \in C. \tag{8.1}$$

In the proof of Theorem 5.1 we used the following fact.

LEMMA 8.1. *Let $g, h : E \to (0, +\infty]$ both be lower semicontinuous. Then*

$$f : \mathbb{R} \to (-\infty, +\infty], \qquad f(y) := \inf\big\{g(x)\big|x \in E \text{ s.t. } h(x) \leq y\big\}$$

is nonincreasing and lower semicontinuous.

Proof. That f is nonincreasing immediately follows from its definition. In order to prove semicontinuity, assume we are given a sequence $(y_n) \subset (-\infty, +\infty]$ converging to some $y \in [-\infty, +\infty]$. We want to show that

$$\liminf_n f(y_n) \geq f(y). \tag{8.2}$$

Observe that with no loss of generality we can restrict ourselves to the case when $y_n \geq \min\{h(x) \,|\, x \in E\}$, since otherwise the set $\{x \in E \text{ s.t. } h(x) \leq y_n\}$ is empty and $f(y_n) = +\infty$. Since h is lower semicontinuous we have that the sets $\{x \in E \text{ s.t. } h(x) \leq y_n\}$ are closed in E. Therefore, since the function g is lower semicontinuous as well, from (8.1) we have that there exists x_n in E such that $f(y_n) = g(x_n)$ and $h(x_n) \leq y_n$. Since the space E is compact, from the sequence (x_n) we can extract a subsequence (x_{n_k}) converging to some $\overline{x}$ in E. From the lower semicontinuity of h we get

$$h(\overline{x}) \leq \liminf_k h\,(x_{n_k}) \leq \liminf_k y_{n_k} = y.$$

It immediately follows that $g(\overline{x}) \geq f(y)$. Finally, from the lower semicontinuity of g we get

$$\liminf_n f(y_n) = \liminf_k g\,(x_{n_k}) \geq g(\overline{x}),$$

which, together with the previous inequality, implies (8.2). ∎

8.2. Proofs for section 4.4. Recall that the interconnection group for the F-labelled ensemble is $S_{Nc} \ltimes F^{Nc}$. We will write the random variable $\Pi_N = (\Pi'_N, \Lambda)$, where Π'_N is uniformly distributed over S_{Nc} and Λ is uniformly distributed over F^{Nc}. For all $s = 1, \ldots, N$, and $k \in G$, let e_s^k in G^N be the vector whose components are all zero but for the sth, which is equal to k.

8.2.1. Proof of Proposition 4.7. Let k in $G \setminus \{0\}$ be such that $a(F,c) = 1 - c + b(Fk,c)$, and define the event $E_s^N := \{e_s^k \in \ker \Phi_N\}$. We have $W_N(\boldsymbol{\tau}_k) = \sum_{s=1}^N \mathbb{1}_{\ker \Phi_N}(e_s^k) = \sum_{s=1}^N \mathbb{1}_{E_s^N}$.

For $1 \leq t \leq L$, define the random variable $N_t := |\Pi'_N(I_s^c) \cap I_t^d|$. Define the event

$$\tilde{E}_s^N := \bigcap_{1 \leq t \leq L} \{N_t = 0\} \cup \{N_t > 0 \text{ and } \exists \text{ closed walk of length } N_t \text{ in } \mathcal{G}(G, Fk)\}.$$

It is not hard to check that $\tilde{E}_s^N \supseteq E_s^N$. Moreover, $\mathbb{P}(E_s^N | \tilde{E}_s^N) \geq |F|^{-c}$, since, given $\tilde{E}_s^N$, there exists at least one realization of the c entries $\Lambda_{(s-1)c+1}, \ldots, \Lambda_{sc}$ in F such that $\Phi_N e_s^k = \mathbf{0}$.

Observe that $\Pi_N(I_s^c)$ is uniformly distributed over the class of all subsets of $\{1, \ldots, Nc\}$ of cardinality c and that there exist at least $\binom{L}{b(Fk,c)}$ possible realizations of $\Pi_N(I_s^c)$ such that, for all $1 \leq t \leq L$, N_t is either 0 or equals the length of a closed walk in $\mathcal{G}(G, Fk)$. It follows that

$$\mathbb{P}(E_s^N) \geq \frac{1}{|F|^c} \mathbb{P}\left(\tilde{E}_s^N\right) \geq \frac{1}{|F|^c} \binom{Nc}{c}^{-1} \binom{L}{b(Fk,c)} \geq K' N^{b(Fk,c)-c} \tag{8.3}$$

for some $K' > 0$ independent of N.

We now estimate the probability of the intersections $E_s^N \cap E_r^N$ for $1 \leq r \neq s \leq N$. We have that, given that E_r^N occurred, $\Pi'_N(I_s^c)$ is uniformly distributed over the class of subsets of of cardinality c of $\{1, \ldots, Nc\} \setminus \Pi'_N(I_r^c)$. It follows that

(8.4)

$$\mathbb{P}(E_s^N | E_r^N) \leq \mathbb{P}(\tilde{E}_s^N | E_r^N) \leq \binom{(N-1)c}{c}^{-1} \binom{L}{b(Fk,c)} \binom{b(Fk,c)d}{c} \leq K'' N^{b(Fk,c)-c}$$

for some $K'' > 0$ independent of N. By applying a union-intersection bound, and using (8.3) and (8.4), we get

$$\begin{aligned} \mathbb{P}\left(W_N(\boldsymbol{\tau}_k) \geq 1\right) &\geq \sum_s \mathbb{P}\left(E_s^N\right) - \sum_{r \neq s} \mathbb{P}\left(E_s^N \cap E_r^N\right) \\ &\geq K' N^{a(F,c)} - K'' N^{2a(F,c)} \geq K N^{a(Fk,c)}, \end{aligned}$$

the last equality holding true for some constant $K > 0$ and N large enough, since $a(F,c) < 0$. ∎

8.2.2. Proof of Proposition 4.8. For $1 \leq s \neq r \leq N$ and $1 \leq t \leq L$, define the event

$$E_{r,s}^N := \bigcap_{t=1}^L \left\{ \left|\Pi_N(I_r^c) \cap I_t^d\right| = \left|\Pi_N(I_s^c) \cap I_t^d\right| \right\}.$$

In the unlabelled (c,d)-regular ensemble $E_{r,s}^N$ is sufficient for the N-tuple $e_r^k - e_s^k$ (whose G-type is $\hat{\boldsymbol{\tau}}_k$) to be in $\ker \Phi_N$. Indeed, in this case each check ends up summing an equal amount of entries equal to k and $-k$. For the F-labelled ensemble it is easy to see that $\mathbb{P}\left(e_r^k - e_s^k \in \ker \Phi_N \middle| E_{r,s}^N\right) \geq |F|^{-2c}$, since, given that $E_{r,s}^N$ occurred, for $\Phi_N(e_r^k - e_s^k)$ to be $\mathbf{0}$ it is sufficient that the $2c$ corresponding labels equal the identity automorphism. Thus,

$$\mathbb{P}\left(W_N\left(\hat{\boldsymbol{\tau}}_k\right) \geq 1\right) \geq \mathbb{P}\left(\sum_{s>r} \mathbb{1}_{\ker \Phi_N}(e_r^k - e_s^k) \geq 1\right) \geq |F|^{-2c} \mathbb{P}\left(\bigcup_{s>r}^N E_{r,s}^N\right).$$

Now we introduce the events $F_r^N := \bigcup_{t=1}^L \left\{ \left|\Pi_N(I_r^c) \cap I_t^d\right| > \frac{d}{2} \right\}$. We have

$$\mathbb{P}\left(F_r^N\right) \leq L \sum_{a=\lfloor d/2 \rfloor + 1}^{c} \binom{c}{a}\binom{d}{a}\binom{dL}{a}^{-1} \leq AN^{-\lfloor d/2 \rfloor}$$

for some positive A independent of N and r. Clearly, we have that F_r^N implies $\overline{E_{r,s}^N}$, so that $\mathbb{P}\left(E_{r,s}^N \middle| F_r^N\right) = 0$. Instead, we have $\mathbb{P}(E_{r,s}^N | \overline{F_r^N}) \geq \binom{(N-1)c}{c}^{-1} \geq (cN)^{-c}$. Thus, there exist some positive N_0 and K' such that, for every $N \geq N_0$,

$$\mathbb{P}\left(E_{r,s}^N\right) \geq \mathbb{P}\left(E_{r,s}^N \middle| \overline{F_r^N}\right)\mathbb{P}\left(\overline{F_r^N}\right) \geq (cN)^{-c}\left(1 - AN^{-\lfloor d/2 \rfloor}\right) \geq K'N^{-c}.$$

For every unordered triple $\{q,r,s\} \subseteq \{1,\dots,N\}$ we consider the event

$$E_{q,r,s}^N := \bigcap_{t=1}^L \left\{ \left|\Pi_N(I_q^c) \cap I_t^d\right| = \left|\Pi_N(I_r^c) \cap I_t^d\right| = \left|\Pi_N(I_s^c) \cap I_t^d\right| \right\}.$$

We have that

$$\mathbb{P}\left(E_{q,r,s}^N\right) \leq (d-1)^c c! \binom{(N-1)c}{c}^{-1} (d-2)^c c! \binom{(N-2)c}{c}^{-1} \leq K''N^{-2c}$$

for some positive K'' independent of N. For every unordered 4-tuple $\{p,q,r,s\}$ define

$$E_{p,q,r,s}^N := \bigcap_{t=1}^L \left\{ \left|\Pi_N(I_p^c) \cap I_t^d\right| = \left|\Pi_N(I_q^c) \cap I_t^d\right| = \left|\Pi_N(I_r^c) \cap I_t^d\right| = \left|\Pi_N(I_s^c) \cap I_t^d\right| \right\}.$$

We have that

$$\mathbb{P}\left(E_{p,q,r,s}^N\right) \leq (d-1)^c c! \binom{(N-1)c}{c}^{-1} (d-2)^c c! \binom{(N-2)c}{c}^{-1} (d-3)^c c! \binom{(N-3)c}{c}^{-1} \leq K'''N^{-3c}$$

for some positive K'' independent of N. It follows that

$$\begin{aligned}
\mathbb{P}\left(W_N\left(\hat{\boldsymbol{\tau}}_k\right) \geq 1\right) &\geq |F|^{-2c}\mathbb{P}\left(\bigcup_{s>r} E_{r,s}^N\right) \\
&\geq \sum_{r<s}\mathbb{P}\left(E_{r,s}^N\right) - \sum_{q<r<s}\mathbb{P}\left(E_{q,r,s}^N\right) - \sum_{p<q<r<s}\mathbb{P}\left(E_{p,q,r,s}^N\right) \\
&\geq \tbinom{N}{2}K'N^{-c} - \tbinom{N}{3}K''N^{-2c} - \tbinom{N}{4}K'''N^{-3c} \\
&\geq KN^{2-c}
\end{aligned}$$

for some positive K independent of N and $N \in \mathcal{N}_{(c,d)}$ large enough. ∎

8.3. Proof of Theorem 5.2. In order to show the first part of the claim, one follows the steps of the proof of Theorem 5.1 until obtaining (5.3) and (5.4). Then (5.3) implies that $\lim_N \mathbb{P}\left(\kappa_N'' < \gamma_{(F,c,d)}\right) = 0$, while from (5.4), since $a(F,c) \leq -1$, one gets $\lim_N \mathbb{P}\left(\kappa_N' < \gamma_{(F,c,d)}\right) \leq KN^{a(F,c)} = 0$.

For the second part of the claim, we first show that

$$\mathbb{P}\left(\liminf_N \mathrm{d}_{\min}\left(\ker \Phi_N\right) \leq \zeta_{(F,c)}\right) = 1. \tag{8.5}$$

Indeed, let us first consider the case $a(F,c) = -1 > 2 - c$. From Proposition 4.7 it follows that, for every $k \in G \setminus \{0\}$ such that $b(Fk,c) = a(F,c) - 1 + c = c - 2$,

$$\sum_{N \in \mathcal{N}_{(c,d)}} \mathbb{P}(W_N(\boldsymbol{\tau}_k) \geq 1) \geq \sum_{N \in \mathcal{N}_{(c,d)}} KN^{a(F,c)} = K \sum_{N \in \mathcal{N}_{(c,d)}} N^{-1} = +\infty.$$

We now recall that by assumption (Π_N) is a sequence of independent random variables, so that the events $\{W_N(\hat{\boldsymbol{\tau}}_k) \geq 1\}$, for N in $\mathcal{N}_{(c,d)}$, are independent. We can thus apply the converse part of the Borel–Cantelli lemma [7] to conclude that with probability one the event $\{W_N(\hat{\boldsymbol{\tau}}_k) \geq 1\}$ occurs for infinitely many $N \in \mathcal{N}_{(c,d)}$. It follows that, for all $K \in G \setminus \{0\}$ such that $b(Fk,c) = c - 2$,

$$\mathbb{P}\big(\liminf_N \mathrm{d}_{\min}(\ker \Phi_N) \leq \boldsymbol{\delta}(k)\big) \geq \mathbb{P}\big(\{W_N(\hat{\boldsymbol{\tau}}_k) \geq 1\} \text{ i. o. } N \in \mathcal{N}_{(c,d)}\big) = 1, \tag{8.6}$$

so that (8.5) follows. The case when $c = 3$ can be treated similarly using Propositions 4.7 and 4.8 and the converse part of the Borel–Cantelli lemma.

It remains to prove that $\liminf_N \mathrm{d}_{\min}(\ker \Phi_N) \geq \zeta_{(F,c)}$ with probability one. First, consider the case $c = 3$. For every k such that $b(Fk,c) = 0$ we have $W_N(\boldsymbol{\tau}_k) = 0$ for every realization of Π_N in the interconnection group $S_{Nc} \ltimes F^{Nc}$. It follows that deterministically

$$\mathrm{d}_{\min}(\ker \Phi_N) \geq \min\big\{(2 - \mathbb{1}_{\{1\}}(b(Fk,c)))\boldsymbol{\delta}(k) \big| k \in G \setminus \{0\}\big\} = \zeta_{(F,c)}.$$

When $c \geq 4$, for every k in $G \setminus \{0\}$ such that $b(Fk,c) < 2 - c$, Lemma 4.5 and the Borel–Cantelli lemma imply that with probability one $\{W_N(\boldsymbol{\tau}_k) = 0\}$ occurs only finitely often. Then using an argument similar to that in the proof of Proposition 4.4 it is possible to show that $\sum_{\frac{1}{N} < ||\boldsymbol{\theta} - \delta_0|| < \frac{2}{d}} \overline{W_N(\boldsymbol{\theta})} \leq KN^{-2}$, and then $\sum_{\frac{1}{N} < ||\boldsymbol{\theta} - \delta_0|| < \frac{2}{d}} W_N(\boldsymbol{\theta}) = 0$ for all but a finitely many N. This implies (8.5). ∎

Acknowledgments. Part of the work was done while the first author was visiting Yale University. We thank the Electrical Engineering Department and Professor Sekhar Tatikonda for their hospitality.

REFERENCES

[1] M. A. Armand, *Decoding LDPC codes over integer residue rings*, IEEE Trans. Inform. Theory, 52 (2006), pp. 4680–4686.

[2] A. Barg and G. D. Forney, Jr., *Random codes: Minimum distances and error exponents*, IEEE Trans. Inform. Theory, 48 (2002), pp. 2568–2573.

[3] S. Benedetto, R. Garello, M. Mondin, and G. Montorsi, *Geometrically uniform TCM codes over groups based on $L \times$ MPSK constellations*, IEEE Trans. Inform. Theory, 40 (1994), pp. 137–152.

[4] A. Bennatan and D. Burshtein, *On the application of LDPC codes to arbitrary discrete memoryless channels*, IEEE Trans. Inform. Theory, 50 (2004), pp. 417–438.

[5] A. Bennatan and D. Burshtein, *Design and analysis of nonbinary LDPC codes for arbitrary discrete memoryless channels*, IEEE Trans. Inform. Theory, 52 (2006), pp. 549–583.

[6] R. Blahut, *Composition bounds for channel block codes*, IEEE Trans. Inform. Theory, 23 (1977), pp. 656–674.

[7] V. S. Borkar, *Probability Theory: An Advanced Course*, Springer, New York, 1995.

[8] J. J. Boutros, A. Ghaith, and Y.-W. Yi, *Non-binary adaptive LDPC codes for frequency selective channels: Code construction and iterative decoding*, in Proceedings of the IEEE Information Theory Workshop, Chengdu, China, 2006, pp. 184–188.

[9] D. Burshetein and U. Miller, *Asymptotic enumeration methods for analyzing LDPC codes*, IEEE Trans. Inform. Theory, 50 (2004), pp. 1115–1131.

[10] G. Como and F. Fagnani, *Ensembles of codes over Abelian groups*, in Proceedings of the IEEE International Symposium on Information Theory, Adelaide, Australia, 2005, pp. 1788–1792.

[11] G. Como and F. Fagnani, *The capacity of finite Abelian group codes over memoryless symmetric channels*, IEEE Trans. Inform. Theory, submitted.
[12] G. Como and F. Fagnani, *On the Gilbert-Varshamov distance of Abelian group codes*, in Proceedings of the IEEE International Symposium on Information Theory, Nice, France, 2007, pp. 2651–2655.
[13] M. C. Davey and D. J. C. MacKay, *Low density parity check codes over GF(q)*, IEEE Comm. Lett., 2 (1998), pp. 159–166.
[14] A. Dembo and A. Montanari, *Finite size scaling for the core of large random hypergraphs*, Ann. Appl. Probab., to appear.
[15] C. Di, T. J. Richardson, and R. Urbanke, *Weight distribution of low-density parity-check codes*, IEEE Trans. Inform. Theory, 52 (2006), pp. 4839–4855.
[16] R. L. Dobrushin, *Asymptotic optimality of group and systematic codes for some channels*, Theor. Probab. Appl., 8 (1963), pp. 47–59.
[17] U. Erez and G. Miller, *The ML decoding performance of LDPC ensembles over $\mathbb{Z}_q$*, IEEE Trans. Inform. Theory, 51 (2005), pp. 1871–1879.
[18] G. D. Forney, Jr., *Geometrically uniform codes*, IEEE Trans. Inform. Theory, 37 (1991), pp. 1241–1260.
[19] R. G. Gallager, *Low Density Parity Check Codes*, MIT Press, Cambridge MA, 1963.
[20] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.
[21] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 5th ed., Oxford University Press, New York, 1979.
[22] J. Hou, P. H. Siegel, L. B. Milstein, and H. D. Pfister, *Capacity-approaching bandwidth-efficient coded modulation schemes based on low-density parity-check codes*, IEEE Trans. Inform. Theory, 49 (2003), pp. 2141–2155.
[23] T. W. Hungerford, *Algebra*, Springer-Verlag, New York, 1974.
[24] R. Koetter, W.-C. W. Li, P. O. Vontobel, and J. L. Walker, *Characterizations of pseudo-codewords of (low-density) parity-check codes*, Adv. Math., 213 (2007), pp. 205–229.
[25] S.-L. Litsyn and V. Shevelev, *On ensembles of low-density parity-check codes: Asymptotic distance distributions*, IEEE Trans. Inform. Theory, 48 (2002), pp. 887–908.
[26] S.-L. Litsyn and V. Shevelev, *Distance distributions in ensembles of irregular low-density parity-check codes*, IEEE Trans. Inform. Theory, 49 (2003), pp. 3140–3159.
[27] H.-A. Loeliger, *Signal sets matched to groups*, IEEE Trans. Inform. Theory, 37 (1991), pp. 1675–1679.
[28] D. J. C. MacKay, *Good error correcting codes based on very sparse matrices*, IEEE Trans. Inform. Theory, 45 (1999), pp. 399–431.
[29] G. Miller and D. Burshetein, *Bounds on the maximum likelihood decoding error probability of low-density parity-check codes*, IEEE Trans. Inform. Theory, 47 (2001), pp. 2696–2710.
[30] K. S. Ng and M. A. Armand, *LDPC codes over mixed alphabets*, Electron. Lett., 42 (2006), pp. 1290–1291.
[31] A. Orlitsky, K. Viswanathan, and J. Zhang, *Stopping set distribution of LDPC code ensembles*, IEEE Trans. Inform. Theory, 51 (2005), pp. 929–953.
[32] V. Rathi and R. Urbanke, *Density evolution, thresholds and the stability condition for non-binary LDPC codes*, IEE Proc. Commun., 152 (2005), pp. 1069–1074.
[33] V. Rathi, *On the asymptotic weight and stopping set distribution of regular LDPC ensembles*, IEEE Trans. Inform. Theory, 52 (2006), pp. 4212–4218.
[34] T. J. Richardson and R. Urbanke, *The capacity of low-density parity-check codes under message-passing decoding*, IEEE Trans. Inform. Theory, 47 (2001), pp. 599–618.
[35] T. J. Richardson, M. A. Shokrollahi, and R. Urbanke, *Design of capacity-approaching irregular low-density parity-check codes*, IEEE Trans. Inform. Theory, 47 (2001), pp. 619–637.
[36] T. J. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, Cambridge, UK, 2007.
[37] I. Sason and R. Urbanke, *Parity-check density versus performance of binary linear block codes over memoryless symmetric channels*, IEEE Trans. Inform. Theory, 49 (2003), pp. 1611–1635.
[38] L. Sassatelli and D. Declercq, *Non-binary hybrid LDPC codes: Structure, decoding and optimization*, in Proceedings of the IEEE Information Theory Workshop, Chengdu, China, 2006, pp. 71–75.
[39] N. Shulman and M. Feder, *Random coding techniques for nonrandom codes*, IEEE Trans. Inform. Theory, 45 (1999), pp. 2001–2004.
[40] D. Sridhara and T. E. Fuja, *LDPC codes over rings for PSK modulation*, IEEE Trans. Inform. Theory, 51 (2005), pp. 3209–3220.

[41] A. Terras, *Fourier Analysis on Finite Groups and Applications*, Cambridge University Press, Cambridge, UK, 1999.
[42] Various Authors, *Special issue on iterative decoding*, IEEE Trans. Inform. Theory, 47 (2001).
[43] P. O. Vontobel and R. Koetter, *Graph-cover decoding and finite-length analysis of message-passing iterative decoding of LDPC codes*, IEEE Trans. Inform. Theory, to appear.
[44] C. C. Wang, S. R. Kulkarni, and H. V. Poor, *Finite-dimensional bounds on $\mathbb{Z}_m$ and binary LDPC codes with belief propagation decoders*, IEEE Trans. Inform. Theory, 53 (2007), pp. 56–81.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 54–58

THE LINEAR ARBORICITY OF GRAPHS ON SURFACES OF NEGATIVE EULER CHARACTERISTIC*

JIAN-LIANG WU†

Abstract. The linear arboricity of a graph G is the minimum number of linear forests which partition the edges of G. In the present, it is proved that if a graph G can be embedded in a surface of Euler characteristic $\varepsilon < 0$ and $\Delta(G) \geq \sqrt{46 - 54\varepsilon} + 19$, then its linear arboricity is $\lceil \frac{\Delta(G)}{2} \rceil$. Some related results on the girth and maximum average degree are also obtained.

Key words. graph, surface, Euler characteristic, linear arboricity

AMS subject classification. 05C05

DOI. 10.1137/S0895480101394690

1. Introduction. In this paper, all graphs are finite, simple, and undirected. Any undefined notation follows that of Bondy and Murty [6]. For a real number x, $\lceil x \rceil$ is the smallest integer not smaller than x, and $\lfloor x \rfloor$ is the largest integer not larger than x. Given a graph $G = (V, E)$, let $N(v) = \{u \mid uv \in E(G)\}$ and $N_k(v) = \{u \mid u \in N(v) \text{ and } d(u) = k\}$, where $d(v) = |N(v)|$ is the *degree* of the vertex v. We use $\Delta(G)$ and $\delta(G)$ to denote the maximum (vertex) degree and the minimum (vertex) degree, respectively. A *k-vertex* is a vertex of degree k. If $W \subseteq V(G)$, then let $N(W) = \bigcup_{v \in W} N(v)$. The *girth* of a graph G is the length of a shortest cycle in G. The *maximum average degree*, denoted by $mad(G)$, of a graph G is the maximum value of $2|E(H)|/|V(H)|$ taken over all subgraphs H of G.

A *linear forest* is a graph in which each component is a path. A map φ from $E(G)$ to $\{1, 2, \ldots, t\}$ is called a *t-linear coloring* if $(V(G), \varphi^{-1}(\alpha))$ is a linear forest for $1 \leq \alpha \leq t$. The *linear arboricity* $la(G)$ of a graph G defined by Harary [10] is the minimum number t for which G has a t-linear coloring. Given a t-linear coloring φ and a vertex v of G, let $C^i_\varphi(v) = \{j \mid \text{the color } j \text{ appears } i \text{ times at } v\}$, where $i = 0, 1, 2$. Then $|C^0_\varphi(v)| + |C^1_\varphi(v)| + |C^2_\varphi(v)| = t$.

Akiyama, Exoo, and Harary [2] conjectured that $la(G) = \lceil (\Delta(G) + 1)/2 \rceil$ for any regular graph G. It is obvious that $la(G) \geq \lceil \Delta(G)/2 \rceil$ for any graph G and $la(G) \geq \lceil (\Delta(G) + 1)/2 \rceil$ for every regular graph G. So the conjecture is equivalent to the following conjecture.

CONJECTURE A. For any graph G, $\lceil \frac{\Delta(G)}{2} \rceil \leq la(G) \leq \lceil \frac{\Delta(G)+1}{2} \rceil$.

The linear arboricity has been determined for complete bipartite graphs [2], Halin graphs [12], series-parallel graphs [14], complete regular multipartite graphs [15], and regular graphs with $\Delta = 3$, 4 [2] and [3], 5, 6, 8 [8], and 10 [9]. Péroche [11] proved that the determination of $la(G)$ of a graph G is a **NP**-hard problem, even when $\Delta = 4$. Alon, Teague, and Wormald [5] proved that there is an absolute constant $c > 0$ such that for every d-regular graph G, $la(G) \leq \frac{d}{2} + cd^{2/3}(\log d)^{1/3}$. A slightly weaker result has been proved in [4, p. 64]. Aït-djafer [1] obtained some results for graphs with multiple edges. For planar graphs, Conjecture A has already been proved to be true;

*Received by the editors September 1, 2001; accepted for publication (in revised form) June 13, 2008; published electronically October 24, 2008. This work was partially supported by National Natural Science Foundation of China (10631070, 60673059).
http://www.siam.org/journals/sidma/23-1/39469.html
†School of Mathematics, Shandong University, Jinan, 250100, People's Republic of China (jlwu@sdu.edu.cn).

see [13] and [17]. Wu also proved in [13] that if a planar graph G has $\Delta \geq 13$, then $la(G) = \lceil \Delta/2 \rceil$, and some related results on the girth are obtained, too. It is noted in [16] that these results can be extended to graphs embeddable in a surface of Euler characteristic $\varepsilon \geq 0$.

In the present paper, we prove that if a graph G is embeddable in a surface of Euler characteristic $\varepsilon < 0$ and $\Delta(G) \geq \sqrt{46 - 54\varepsilon} + 19$, then $la(G) = \lceil \Delta/2 \rceil$. We also consider the relationship between linear arboricity and $mad(G)$. Here surfaces are all compact, connected 2-manifolds with boundary and any embedding of graphs are 2-cell embedding.

2. Main results and their proofs. First, let us describe a result proved by Borodin, Kostochka, and Woodall [7]. Let G be a graph, and let $f : E(G) \to \mathbb{N}$ be a function into the positive integers. A *proper edge coloring* of G is a coloring of $E(G)$ such that no two adjacent edges receive the same color. A graph G is said to be *edge f-choosable* if, whenever we give lists of $f(e)$ colors to each edge e of G, there exists a proper edge coloring of G where each edge is colored with a color from its own list.

LEMMA 2.1 (see [7]). *A bipartite graph G is edge f-choosable where $f(e) = \max\{d(u), d(v)\}$ for $e = uv \in E(G)$.*

If φ is a t-linear coloring of a graph G, a vertex $v \in V(G)$, and $i \in \{0, 1, 2\}$, then let $C^i_\varphi(v) = \{j \mid \text{the color } j \text{ appears } i \text{ times at } v\}$. Then $|C^0_\varphi(v)|+|C^1_\varphi(v)|+|C^2_\varphi(v)| = t$ and $|C^1_\varphi(v)| + 2|C^2_\varphi(v)| = d(v)$, so that

$$2|C^0_\varphi(v)| + |C^1_\varphi(v)| = 2t - d(v). \tag{1}$$

We now state and prove our main result.

THEOREM 2.2. *Let $d \geq \sqrt{46 - 54\varepsilon} + 19$, and let G be a graph with maximum degree $\Delta(G) \leq d$, embedded in a surface of Euler characteristic $\varepsilon < 0$. Then* $\mathrm{la}(G) \leq \lceil \frac{d}{2} \rceil$. *In particular, if $\Delta(G) = d$, then* $\mathrm{la}(G) = \lceil \frac{\Delta(G)}{2} \rceil$.

Proof. Let G be a minimum counterexample to the theorem. First, we prove some claims for G.

Claim 1. For any $uv \in E(G)$, $d_G(u) + d_G(v) \geq d + 2$.

Proof of Claim 1. Suppose that G has an edge uv with $d_G(u) + d_G(v) \leq d + 1$. Then $G' = G - uv$ has a $\lceil \frac{d}{2} \rceil$-linear coloring φ by the minimality of G. Let $S = C^2_\varphi(u) \cup C^2_\varphi(v) \cup (C^1_\varphi(u) \cap C^1_\varphi(v))$. Since $d_{G'}(u) + d_{G'}(v) = d(u) + d(v) - 2 \leq d - 1$, $|S| < \lceil \frac{d}{2} \rceil$. Let $\varphi(uv) \in \{1, 2, \ldots, \lceil \frac{d}{2} \rceil\} \backslash S$. Thus φ is extended to a $\lceil \frac{d}{2} \rceil$-linear coloring of G, a contradiction. Hence Claim 1 holds. ∎

By Claim 1, we have $\delta(G) \geq 2$ and any two 2-vertices are not adjacent.

Claim 2. G has no even cycle $v_0v_1 \cdots v_{2n-1}v_0$ such that $d(v_1) = d(v_3) = \cdots = d(v_{2n-1}) = 2$ and $\max_{0 \leq i < n} |N_2(v_{2i})| \geq 3$.

Proof of Claim 2. Suppose G does contain such an even cycle. Without loss of generality, let $N_2(v_0) \geq 3$, which implies that v_0 is adjacent to at least three 2-vertices. Let $u \in N_2(v_0) \backslash \{v_{2n-1}, v_1\}$ and $v \in N(u) \backslash v_0$. By the induction hypothesis, $G^* = G - \{v_1, \ldots, v_{2n-1}\} - uv_0$ has a $\lceil \frac{d}{2} \rceil$-linear coloring φ. Now we construct directly a $\lceil \frac{d}{2} \rceil$-linear coloring σ of G as follows.

First of all, if $C^0_\varphi(v_0) \neq \emptyset$, let $\sigma(uv_0) = \sigma(v_0v_1) \in C^0_\varphi(v_0)$. Otherwise, $|C^1_\varphi(v_0)| \geq 3$, let $\sigma(uv_0) \in C^1_\varphi(v_0) \backslash \varphi(uv)$ and $\sigma(v_1v_0) \in C^1_\varphi(v_0) \backslash \sigma(uv_0)$. After that, let $\sigma(v_0v_{2n-1}) \in (C^1_\varphi(v_0) \cup C^0_\varphi(v_0)) \backslash \{\sigma(uv_0), \sigma(v_0v_1)\}$. So $\sigma(v_0v_1) \neq \sigma(v_0v_{2n-1})$. Furthermore, for $i = 1, 2, \ldots, n-1$, if $\sigma(v_0v_{2n-1}) \in C^1_\varphi(v_{2i})$, let $\sigma(v_{2i-1}v_{2i}) = \sigma(v_0v_{2n-1})$. Otherwise, let $\sigma(v_{2i-1}v_{2i}) \in (C^1_\varphi(v_{2i}) \backslash \sigma(v_{2i-2}v_{2i-1})) \cup C^0_\varphi(v_{2i})$. And $\sigma(v_{2i}v_{2i+1}) \in$

$(C^1_\varphi(v_{2i})\backslash\sigma(v_{2i-1}v_{2i})) \cup C^0_\varphi(v_{2i})$. Finally, the uncolored edges of G are colored the same colors as in φ of G^*. This contradiction proves Claim 2. □

Let G_2 be the subgraph induced by edges incident with 2-vertices. Since G does not contain two adjacent 2-vertices, G_2 does not contain any odd cycle. So it follows from Claim 2 that any component of G_2 is either an even cycle or a tree. So it is easy to find a matching M in G saturating all 2-vertices (M contains alternate edges of every even cycle of G_2, and if some component of G_2 is a tree T, then we repeatedly add to M a pendant edge e of T and delete the endvertices of e from T). If $uv \in M$ and $d(u) = 2$, then v is called a 2-*master* of u. Note that every 2-vertex has a 2-master, which is necessarily a vertex of maximum degree, and each vertex of the maximum degree can be the 2-master of at most one 2-vertex.

For an integer t $(3 \le t \le \lfloor \frac{d}{2} \rfloor)$, let $X_t \subseteq \{v \mid 2 \le d_G(v) \le t\}$ and $Y_t = N(X_t)$. It follows from Claim 1 that X_t is an independent set of G. Let K be the induced bipartite subgraph of G with partite sets X_t and Y_t. Then $d_K(u) = d_G(u)$ for $u \in X_t$. If $d_K(v) \ge d_G(v) + 2(t - \lceil \frac{d}{2} \rceil)$ for each $v \in Y_t$, then K is called *t-alternating*.

Claim 3. G contains no t-alternating subgraph for any t $(3 \le t \le \lfloor \frac{d}{2} \rfloor)$.

Proof of Claim 3. Suppose that for some t $(3 \le t \le \lfloor \frac{d}{2} \rfloor)$, G contains a t-alternating subgraph H with partite sets X and Y such that $2 \le d_H(x) = d_G(x) \le t$ for $x \in X$ and $d_H(y) \ge 2t + d_G(y) - 2\lceil \frac{d}{2} \rceil$ for each $v \in Y$. Then there is a $\lceil \frac{d}{2} \rceil$-linear coloring σ to color all edges of $G - X$ by the minimality of G.

Let $F = (X, Y')$ be the bipartite graph obtained from H by splitting equitably each vertex $v \in Y$ into two vertices v_1 and v_2, that is, such that $v_1, v_2 \in Y'$ and $\lfloor \frac{d_H(v)}{2} \rfloor = d_F(v_1) \le d_F(v_2) = \lceil \frac{d_H(v)}{2} \rceil$. Similarly, split equitably the set $C^1_\sigma(v)$ into two subsets C' and C'', that is, $C^1_\sigma(v) = C'_{v_1} \cup C'_{v_2}$ and $\lfloor \frac{C^1_\sigma(v)}{2} \rfloor = |C'_{v_1}| \le |C'_{v_2}| = \lceil \frac{C^1_\sigma(v)}{2} \rceil$. Thus for each vertex $v \in Y$ and its splitting vertices $v_1, v_2 \in Y'$, let $C_{v_1} = C^0_\sigma(v) \cup C'$ and $C_{v_2} = C^0_\sigma(v) \cup C''$. It follows that for any $xy \in E(F)$ with $x \in X$ and $y \in Y'$, $|C_y| \ge \max\{t, d_F(y)\} \ge \max\{d_F(x), d_F(y)\}$ since $|C^0_\sigma(v)| + |C^1_\sigma(v)| + |C^2_\sigma(v)| = \lceil \frac{d}{2} \rceil$ and $2|C^2_\sigma(v)| + |C^1_\sigma(v)| = d_G(v) - d_H(v)$. Now define the list A_{xy} of xy as C_y. By Lemma 2.1, any edge xy of F can be colored from its list A_{xy}. If we use the same coloring to return to color all edges of H, then we extend σ to a $\lceil \frac{d}{2} \rceil$-linear coloring of G, a contradiction with G being a counterexample. So this contradiction proves Claim 3. □

Claim 4. If $X_t \ne \emptyset$, then there exists a bipartite subgraph M_t of K_t such that $d_{M_t}(x) = 1$ for each $x \in X_t$, and $0 \le d_{M_t}(y) \le 2t - 1$ for each $y \in Y_t$.

Proof of Claim 4. Let $H_t = (X'_t, Y_t)$, where $X'_t \subseteq X_t$, be a maximum bipartite subgraph of K_t such that $d_{H_t}(x) = 1$ for $x \in X'_t$ and $d_{H_t}(y) \le 2t - 1$ for $y \in Y_t$. Clearly, H_t is not empty since G has at least one edge from X_t to Y_t. Suppose $X'_t \ne X_t$. Let $v \in X_t \backslash X'_t$. An *alternating path*, P_v, in K_t is a path whose origin is v and edges are alternating between $E(K_t) \setminus E(H_t)$ and $E(H_t)$. If K_t has an alternating path $P_v = vv_1v_2 \cdots v_{2m+1}$ such that its terminus v_{2m+1} is in Y_t and $d_{H_t}(v_{2m+1}) \le 2t - 2$, then $H'_t = (H_t - \{v_1v_2, v_3v_4, \ldots, v_{2m-1}v_{2m}\}) + \{vv_1, v_2v_3, \ldots, v_{2m}v_{2m+1}\}$ is another bipartite subgraph satisfying the claim, but $|E(H'_t)| > |E(H_t)|$, a contradiction to the maximality of H_t. So for every alternating path P_v whose terminus is a vertex $v' \in Y_t$, we have $d_{H_t}(v') = 2t - 1$. Let Z_t denote the set of all vertices connected to v by alternating paths. Set $X''_t = Z_t \cap X_t$ and $Y''_t = Z_t \cap Y_t$. Then $X''_t = \{v\} \cup (Z_t \cap X'_t)$, $Y''_t = N(X''_t)$, and $d_{H_t}(y) = 2t - 1$ for any $y \in Y''_t$. Let F_t be the bipartite subgraph induced by edges between X''_t and Y''_t. Then $d_{F_t}(y) \ge d_{H_t}(y) + 1 = 2t \ge 2t + d_G(y) - 2\lceil \frac{d}{2} \rceil$ for any $y \in Y''_t$. By the definition of X_t, $d_G(x) = d_{F_t}(x) \le t$ for any $x \in X''_t$. So F_t is a t-alternating subgraph of G, a contradiction to Claim 3. Hence $X_t = X'_t$ and

Claim 4 is true. □

Here we call y the *t-master* of x in G for $xy \in M_t$. In particular, it follows from the claim that for each i and j $(2 \le i \le j \le 5)$, every i-vertex has a j-master. We shall use the important idea to redistribute charge below.

By Euler's formula $|V| - |E| + |F| \ge \varepsilon$, and by the fact that $2|E| \ge 3|F|$, we have

$$\sum_{x \in V} (d(x) - 6) = 2|E| - 6|V| \le 6|F| - 4|E| - 6\varepsilon \le -6\varepsilon.$$

Define a charge ω on vertices of G by letting $\omega(v) = d(v) - 6$ for $v \in V(G)$. Now we construct a new charge ω^* from ω by the following rule.

For each i and j $(2 \le i \le j \le 5)$, each i-vertex receives charge 1 from its j-masters. Clearly, $\sum_{v \in V(G)} \omega(v) = \sum_{v \in V(G)} \omega^*(v) \le -6\varepsilon$. We will get a contradiction by proving that $\sum_{v \in V(G)} \omega^*(v) > -6\varepsilon$.

Claim 5. For each vertex $v \in V$, $\omega^*(v) \ge 0$; moreover, $\omega^*(v) \ge \lfloor \frac{d}{3} \rfloor - 8$ if $d(v) \ge \lfloor \frac{d}{3} \rfloor$.

Proof of Claim 5. If $2 \le d(v) \le 5$, then $\omega^*(v) = 0$ since v receives $6 - d(v)$ in total from its j-masters where $j = d(v), d(v) + 1, \ldots, 5$. If $d(v) = 6$, then $\omega^*(v) = 0$. If $7 \le d(v) \le d - 4$, then v receives and sends nothing in the redistribution since $d_G(u) \ge 6$ for each $u \in N(v)$ by Claim 1; so $\omega^*(v) = \omega(v) = d(v) - 6 > 0$, and, moreover, $d(v) - 6 \ge \lfloor \frac{d}{3} \rfloor - 8$ if $d(v) \ge \lfloor \frac{d}{3} \rfloor$. If $d(v) = d - 3$, then neighbors of v have a degree of at least 5. This implies that v may be a 5-master of at most 9 vertices in G by Claim 4. So $\omega^*(v) \ge \omega(v) - 9 = ((d-3) - 9) - 6 = (d - 12) - 6$. If $d(v) = d - 2$, then $d_G(u) \ge 4$ for $u \in N(v)$, and it may be a 5-master of at most 9 vertices and a 4-master of at most 7 vertices. So $\omega^*(v) \ge \omega(v) - 9 - 7 = (d-18) - 6$. Similarly, we have $\omega^*(v) \ge \omega(v) - 9 - 7 - 5 = (d - 22) - 6$ if $d(v) = d - 1$ and $\omega^*(v) \ge \omega(v) - 9 - 7 - 5 - 1 = (d - 22) - 6$ if $d(v) = d$. Hence $\omega^*(v) \ge (d - 20) - 8$ if $d_G(v) \ge d - 3$. Since $d \ge \sqrt{46 - 54\varepsilon} + 19$, $d - 20 \ge \lfloor \frac{d}{3} \rfloor$. So $\omega^*(v) \ge \lfloor \frac{d}{3} \rfloor - 8$ if $d(v) \ge \lfloor \frac{d}{3} \rfloor$. Hence we prove Claim 5. □

Let $U = \{u \mid d_G(u) \le \lfloor \frac{d}{3} \rfloor\}$ and $W = N(U)$. Then U is an independent set of G by Claim 1. Let F be the induced bipartite subgraph of G with partite sets U and W. If $|V(G) \backslash U| \le \lfloor \frac{d}{3} \rfloor + 1$, then for any vertex $w \in W$, $d_F(w) = d_G(w) - d_{G-U}(w) \ge d_G(w) - \lfloor \frac{d}{3} \rfloor \ge d_G(w) - 2\lceil \frac{d}{2} \rceil + 2\lfloor \frac{d}{3} \rfloor$, that is, F is a $(\lfloor \frac{d}{3} \rfloor)$-alternating subgraph of G, a contradiction to Claim 3. So $|V(G) \backslash U| \ge \lfloor \frac{d}{3} \rfloor + 2$. Thus we have $\sum_{v \in V(G)} \omega(v) = \sum_{v \in V(G)} \omega^*(v) \ge (\lfloor \frac{d}{3} \rfloor + 2)(\lfloor \frac{d}{3} \rfloor - 8) \ge (\lfloor \frac{\sqrt{46-54\varepsilon}+19}{3} \rfloor + 2)(\lfloor \frac{\sqrt{46-54\varepsilon}+19}{3} \rfloor - 8) > -6\varepsilon$, a contradiction. This completes the proof. □

If the girth of a graph G embedded in a surface of Euler characteristic $\varepsilon < 0$ is at least 4, then $|E(G)| \le 2(|V(G)| - \varepsilon)$, that is, $\sum_{x \in V} (d(v) - 4) \le -4\varepsilon$. By using a similar argument we can prove the following theorem.

THEOREM 2.3. *Let G be a graph embedded in a surface of Euler characteristic $\varepsilon < 0$. If G has girth at least 4 and $\Delta(G) \ge \sqrt{45 - 36\varepsilon} + 7$, then* $\mathrm{la}(G) = \lceil \frac{\Delta(G)}{2} \rceil$.

We close the paper with a result on the maximum average degree.

THEOREM 2.4. *Let G be a graph with $\mathrm{mad}(G) \le t$ for some integer $t \ge 2$. If $\Delta(G) \ge (t+1)(t-2) + 1$, then* $\mathrm{la}(G) = \lceil \frac{\Delta(G)}{2} \rceil$.

Proof. Let G be a minimal counterexample. Since $t \ge 2$ and $\Delta(G) \ge (t+1)(t-2) + 1$, $t \le \lfloor \frac{\Delta(G)}{2} \rfloor$. Thus it follows from the proof of Theorem 2.2 that $\delta(G) \ge 2$, G has no k-alternating subgraph for any $2 \le k \le t$, G has no even cycle $v_0 v_1 \cdots v_{2n-1} v_0$ such that $d(v_1) = d(v_3) = \cdots = d(v_{2n-1}) = 2$ and $\max_{0 \le i < n} |N_2(v_{2i})| \ge 3$, and

$d_G(u) + d_G(v) \geq \Delta(G) + 2$ if $uv \in E(G)$. So every i-vertex has a j-master for $2 \leq i \leq t-1$ and $j = i, i+1, \ldots, t-1$.

Since $\text{mad}(G) \leq t$, $\sum_{x \in V}(d(x) - t) \leq 0$. Define a charge ω on vertices of G by letting $\omega(v) = d(v) - t$ for $v \in V(G)$. Now we construct a new charge ω^*: Each i-vertex receives 1 from all its j-masters where $2 \leq i < t$ and $j = i, i+1, \ldots, t-1$.

It is obvious that $\omega^*(v) = 0$ if $d(v) = 2, 3, \ldots, t$ and $\omega^*(v) = \omega(v) > 0$ if $t < d(v) \leq \Delta(G) - t + 2$. If $d(v) = \Delta(G) - k$ where $2 \leq k \leq t-3$, then $\omega^*(v) \geq \omega(v) - (2t-3) - (2t-5) - \cdots - (2k-1) > 0$. If $d(v) = \Delta(G) - 1$, then $\omega^*(v) \geq \omega(v) - (2t-3) - (2t-5) - \cdots - 5 > 0$. If $d(v) = \Delta(G)$, then $\omega^*(v) \geq \omega(v) - (2t-3) - (2t-5) - \cdots - 5 - 1 = \Delta(G) - (t+1)(t-2) > 0$. Therefore, $\sum_{v \in V(G)} \omega(v) = \sum_{v \in V(G)} \omega^*(v) > 0$, a contradiction. This completes the proof. □

REFERENCES

[1] H. Aït-djafer, *Linear arboricity for graphs with multiple edges*, J. Graph Theory, 11 (1987), pp. 135–140.

[2] J. Akiyama, G. Exoo, and F. Harary, *Covering and packing in graphs* III: *Cyclic and acyclic invariants*, Math. Slovaca, 30 (1980), pp. 405–417.

[3] J. Akiyama, G. Exoo, and F. Harary, *Covering and packing in graphs* IV: *Linear arboricity*, Networks, 11 (1981), pp. 69–72.

[4] N. Alon and J. H. Spencer, *The Probabilistic Method*, John Wiley & Sons, New York, 1992.

[5] N. Alon, V. J. Teague, and N. C. Wormald, *Linear arboricity and linear k-arboricity of regular graphs*, Graphs Combin., 17 (2001), pp. 11–16.

[6] J. A. Bondy and U. S. R. Murty, *Graph Theory with Applications*, American Elsevier, New York, 1976.

[7] O. V. Borodin, A. V. Kostochka, and D. R. Woodall, *List edge and list total colourings of multigraphs*, J. Combin. Theory Ser. B, 71 (1997), pp. 184–204.

[8] H. Enomoto and B. Peroche, *The linear arboricity of some regular graphs*, J. Graph Theory, 8 (1984), pp. 309–324.

[9] F. Guldan, *The linear arboricity of* 10 *regular graphs*, Math. Slovaca, 36 (1986), pp. 225–228.

[10] F. Harary, *Covering and packing in graphs* I, Ann. New York Acad. Sci., 175 (1970), pp. 198–205.

[11] B. Peroche, *Complexity of the linear arboricity of a graph*, RAIRO Rech. Opér., 16 (1982), pp. 125–129 (in French).

[12] J. L. Wu, *Some path decompositions of Halin graphs*, Shandong Kuangye Xueyuan Xuebao, 17 (1998), pp. 92–96 (in Chinese).

[13] J. L. Wu, *On the linear arboricity of planar graphs*, J. Graph Theory, 31 (1999), pp. 129–134.

[14] J. L. Wu, *The linear arboricity of series-parallel graphs*, Graphs Combin., 16 (2000), pp. 367–372.

[15] J. L. Wu, G. Z. Liu, and Y. L. Wu, *The linear arboricity of composition graphs*, J. Syst. Sci. Complex., 15 (2002), pp. 372–375.

[16] J. L. Wu and L. Y. Miao, *The linear arboricity of graphs*, Adv. Math., 27 (1998), pp. 561–562.

[17] J. L. Wu and Y. W. Wu, *The linear arboricity of planar graphs of maximum degree seven are four*, J. Graph Theory, 58 (2008), pp. 210–220.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 59–78

ON COSETS OF WEIGHT 4 OF $BCH(2^m, 8)$, m EVEN, AND EXPONENTIAL SUMS*

PASCALE CHARPIN†, TOR HELLESETH‡, AND VICTOR ZINOVIEV§

Abstract. We give exact expressions for the number of coset leaders in the cosets of weight 4 of binary primitive narrow sense Bose–Chaudury–Hocquenghem (BCH) codes of length $n = 2^m$ (m even) with minimum distance 8 in terms of several exponential sums, including cubic sums and Kloosterman sums. This allows us to bound the number of coset leaders in these cosets.

Key words. binary primitive narrow sense BCH code, coset, coset weight distribution, exponential sum, cubic sum, Kloosterman sum, partial sum, inverse cubic sum

AMS subject classifications. 11T71, 11T23

DOI. 10.1137/070692649

1. Introduction. This paper is a natural continuation of our previous papers [3], [4], [5], and [6]. In these papers, we studied the coset weight distributions of binary extended triple-error-correcting primitive narrow sense Bose–Chaudury–Hocquenghem (BCH) codes. Such a code is of length 2^m and minimum distance 8, which we will denote by $BCH(2^m, 8)$, and is the extension of the binary cyclic code of length $2^m - 1$ and designed distance 7, i.e., the cyclic code with zeros set $\{\alpha,\ \alpha^3,\ \alpha^5\}$ (where α is a primitive root of the finite field of order 2^m).

In [3] and [4] we described coset weight distributions of $BCH(2^m, 8)$ for odd m for the cosets of any weight $j = 1, 2, 3, 4, 5, 6$. For the cosets of weight 4, using an approach developed in [11], we have found [4] the exact expressions for the number of words of weight 4 in terms of the exponential sums of four different types, in particular, of the Kloosterman sums over $GF(2^m)$. Using these results we obtained new properties of Kloosterman sums, mainly their divisibility modulo 24 (see [5]).

The purpose of this paper is to obtain similar results in the case where m is even. Here we extend these results for even m, obtaining explicit expressions for the number of words of weight 4 of cosets of weight 4 of $BCH(2^m, 8)$. For the codes $BCH(2^m, 8)$ the case of even m is much harder, since the exact expressions depend on five different exponential sums. Analyzing these sums we reduce the final expressions to the exponential sums of four different types, including cubic sums and Kloosterman sums. Known bounds for values of these sums permit us to bound the number of words of weight 4 in the cosets of weight 4.

This paper is organized as follows. In section 2, following [3] and [10] we give some preliminary results concerning the codes $BCH(2^m, 8)$ and exponential sums over $GF(2^m)$, in particular, the cubic sums and Kloosterman sums. In section 3 we consider a nonlinear system of equations, which defines the number of words of weight

*Received by the editors May 22, 2007; accepted for publication (in revised form) June 16, 2008; published electronically October 24, 2008. This work was supported by INRIA-Rocquencourt, by the Norwegian Research Council under grant 171094/V30, and also by the Russian Fund of Fundamental Researches (project 06-01-00226).

http://www.siam.org/journals/sidma/23-1/69264.html

†INRIA, Domaine de Voluceau-Rocquencourt, BP 105-78153, Le Chesnay, France (pascale.charpin@inria.fr).

‡Department of Informatics, University of Bergen, N-5020 Bergen, Norway (torh@ii.uib.no).

§Institute for Problems of Information Transmission, Russian Academy of Sciences, Bol'shoi Karetnyi per. 19, GSP-4, Moscow, 101447, Russia (zinov@iitp.ru).

4 in a coset of weight 4 of a code $BCH(2^m, 8)$. In section 4 we solve the nonlinear system of equations, which gives the number of words of weight 4 for any such coset. We express the number of solutions to this system in terms of the exponential sums of four different types: the two cubic sums, the Kloosterman sums, and the so-called inverse cubic sum. Here we use the same approach as in [10], [11], [12]. Using known results on exponential sums, we lower and upper bound the number of words of weight 4 in any coset of weight 4. In section 5 we compute all of the possible values of the number of words for the first nontrivial values $m = 6$ and $m = 8$.

2. Definitions and preliminary results. The Hamming *weight* of any vector (or word) x is denoted by $wt(x)$. Generally, we denote by $\mathbf{F}_{2^k}$ the finite field of order 2^k. However, we simply denote by $\mathbf{F}$ the field $\mathbf{F}_{2^m}$. For any set E containing 0 we denote: $E^* = E \setminus \{0\}$. Also, $\#E$ denotes the cardinality of any set E.

Let us denote by $BCH(2^m, 8)$ a binary primitive (in narrow sense) extended BCH code of length $n = 2^m$, where $m \geq 5$, and the minimal distance is 8. This is the code over $GF(2)$ with the parity check matrix given by

$$H_B \;=\; \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & \alpha & \alpha^2 & \cdots & \alpha^{n-2} \\ 0 & 1 & \alpha^3 & \alpha^6 & \cdots & \alpha^{(n-2)3} \\ 0 & 1 & \alpha^5 & \alpha^{10} & \cdots & \alpha^{(n-2)5} \end{bmatrix},$$

where α is a primitive root of $\mathbf{F}$ (see [16, ch. 7, section 6]). We use the elements of $\mathbf{F}$ as locators for the code $BCH(2^m, 8)$, where the first position of $BCH(2^m, 8)$ corresponds to the zero element of $\mathbf{F}$.

Let $D = x + BCH(2^m, 8)$ be a coset of $BCH(2^m, 8)$. *The weight of the coset* D is the minimum weight of the words of D. *A leader* of D is a codeword of D of minimum weight. To this coset D we associate a *syndrome*, which is a vector, say S, over $\mathbf{F}$ with four coordinates:

$$S \;=\; (S_1, S_2, S_3, S_4) \;=\; xH_B^t,$$

where x is any vector from D and H_B^t is the transpose of the matrix H_B. In this paper we consider only cosets D of weight four. Since the first component S_1 of the syndrome S shows the parity of the vector x, in the rest of this paper, under a syndrome of a coset D, we use the vector (S_2, S_3, S_4), i.e., without the first (zero) coordinate. Recall that the covering radius of $BCH(2^m, 8)$ is 6 [9]. Therefore, the weight i of D is in the set $\{0, \ldots, 6\}$.

Let $Tr(x)$ denote the absolute trace of $x \in \mathbf{F}$ and, for even m, denote by $x \mapsto Tr_2^m(x)$ the trace function from $\mathbf{F}$ to its subfield $\mathbf{F}_4$.

LEMMA 1 (see [11]). *Let a, b be two arbitrary elements of $\mathbf{F}^*$, $a \neq b$. Then*

$$Tr\left(\frac{a\,b}{(a+b)^2}\right) \;=\; 0.$$

LEMMA 2 (see [15]). *The quadratic equation $x^2 + ax + b = 0$, $a \in \mathbf{F}^*, b \in \mathbf{F}$, has two different roots in $\mathbf{F}$ if $Tr(b/a^2) = 0$ and no roots in $\mathbf{F}$ if $Tr(b/a^2) = 1$.*

LEMMA 3 (see [1]). *The cubic equation $x^3 + ax + b = 0$, where $a \in \mathbf{F}$ and $b \in \mathbf{F}^* = \mathbf{F} \setminus \{0\}$, has a unique solution in $\mathbf{F}$ if and only if $Tr(a^3/b^2) \neq Tr(1)$. Furthermore, if it has three distinct roots in $\mathbf{F}$, then $Tr(a^3/b^2) = Tr(1)$.*

Denote $f_b(x) = x^3 + x + b$, where $b \in \mathbf{F}^*$. Let

$$M_i = \#\{\ b\ :\ f_b(x) \text{ has precisely } i \text{ zeros in } \mathbf{F}\ \}.$$

LEMMA 4 (see [13]). *Let $n = 2^m$ where m is even. Then clearly $M_2 = 0$ and*

$$\begin{aligned} M_0 &= (n-1)/3, \\ M_1 &= n/2, \\ M_3 &= (n-4)/6. \end{aligned}$$

Denote

$$e(a) \ = \ (-1)^{Tr(a)}.$$

The function $e(x)$ is an additive character of $\mathbf{F}$. For any mapping $f : \mathbf{F} \mapsto \mathbf{F}$, the expression of the type

$$\sum_{x \in \mathbf{F}} e(f(x))$$

is called an exponential (or a character) sum over $\mathbf{F}$.

LEMMA 5 (see [4]). *Let σ be any mapping from $\mathbf{F}$ to $\mathbf{F}$, and let $\lambda \in \mathbf{F}^*$. Denote by H the kernel of the linear function $x \mapsto Tr(\lambda x)$. Then*

$$\sum_{x \in F} e(\sigma(x)) + \sum_{x \in F} e(\sigma(x) + \lambda x) = 2 \sum_{x \in H} e(\sigma(x)).$$

The exponential sums of polynomials of degree three over $\mathbf{F}$ are known; they are known also from coding theory (see [16, chapter 15]). In particular, we need the following result due to Carlitz [2]. For arbitrary elements $a \in \mathbf{F}^*$ and $b \in \mathbf{F}$, denote

$$C(a,\, b) \ = \ \sum_{x \in \mathbf{F}} e(a\, x^3 \ + \ b\, x), \quad C(a) \ = \ C(a, 0).$$

LEMMA 6 (see [2]). *Let $a \in \mathbf{F}^*$. For any even $m = 2s$ we have that*

$$C(a) \ = \ \begin{cases} (-1)^{s+1} 2^{s+1} & \text{if } a \text{ is a cube in } \mathbf{F}, \\ (-1)^s 2^s & \text{otherwise.} \end{cases}$$

If $a = \beta^3, \ \ \beta \in \mathbf{F}$, then

$$C(a,b) \ = \ \begin{cases} (-1)^{s+1} 2^{s+1} e(x_0^3) & \text{if } Tr_2^m(b\beta^{-1}) = 0, \\ 0 & \text{otherwise,} \end{cases}$$

where x_0 denotes any solution of $x^4 + x = \beta^{-2} b^2$.

If $a \neq \beta^3, \ \ \beta \in \mathbf{F}$, then

$$C(a,b) \ = \ (-1)^s 2^s\, e(a\, x_1^3),$$

where x_1 is the unique solution of $a^2 x^4 + a\, x \ = \ b^2$, given by

$$\left(a^{(2^{2s}-1)/3} + 1\right)\, x_1 \ = \ \sum_{j=0}^{s-1} \left(a^{-1} b^2\right)^{2^{2j}}\ a^{(2^{2j}-1)/3}.$$

We also need the exponential sums of such type for the case when the argument x runs over $\mathbf{F}$ with the fixed trace of the element $1/x$. It is convenient for us to define this partial sum multiplied by 2:

$$P(a,\, b) \;=\; 2 \sum_{x\in\mathbf{F}:\; Tr(1/x)=0} e(a\,x^3 \;+\; b\,x). \tag{2.1}$$

Recall that the classical binary Kloosterman sum, say $K'(a)$, is defined for each a in $\mathbf{F}^*$ by

$$K'(a) \;=\; \sum_{x\in\mathbf{F}^*} e\left(a\,x \;+\; \frac{1}{x}\right).$$

The exponential sums, which we consider here, are generally defined on $\mathbf{F}^*$, the multiplicative group of $\mathbf{F}$. In this paper we extend all of the sums to 0, assuming that $e(x^{-1}) = e(x^{-3}) = 1$ for $x = 0$. Indeed, $Tr(x^{-1}) = Tr(x^{2^{m-1}-1})$ so that we can define $Tr(x^{-1}) = 0$ for the case $x = 0$. Therefore, we define here the classical Kloosterman sum $K(a),\ a \in \mathbf{F}^*$, as

$$K(a) \;=\; \sum_{x\in\mathbf{F}} e\left(a\,x \;+\; \frac{1}{x}\right) \;=\; K'(a) \;+\; 1. \tag{2.2}$$

We extend the sum $K'(a)$ to $a = 0$, setting $K(0) \;=\; 0$.

Note that we have (where $x = ya$ and $z^2 = y$)

$$\sum_{x\in\mathbf{F}} e\left(\frac{a}{x} + ax\right) = \sum_{y\in\mathbf{F}} e\left(\frac{1}{y} + a^2 y\right) = \sum_{z\in\mathbf{F}} e\left(\frac{1}{z} + az\right) = K(a). \tag{2.3}$$

And obviously $K(a) \;=\; K(a^2)$.

Using deep results on the number of rational points on certain elliptic curves, Lachaud and Wolfmann [14] proved the following result.

LEMMA 7. *The set $K(a),\ a \in \mathbf{F}$ is the set of all the integers $s \equiv\ 0 \pmod 4$ with value s in the range $[-2^{(m/2)+1} + 1,\ 2^{(m/2)+1} + 1]$.*

Note that we deduce immediately that for m even and for any $a \in \mathbf{F}$, we have

$$-2^{(m/2)+1} + 4 \le K(a) \le 2^{(m/2)+1}. \tag{2.4}$$

Considering the coset weight distribution of Z_4-linear Goethals codes, we obtained the following result.

LEMMA 8 (see [10]). *For any $m \ge 3$,*

$$K(a) \;\equiv\; \begin{cases} 4 \mod 8 & if \quad Tr(a) = 1, \\ 0 \mod 8 & if \quad Tr(a) = 0. \end{cases}$$

We also need the following observation, partly given in [4].

LEMMA 9. *For any $a \in \mathbf{F}^*$ and for any m,*

$$K(a) = 2 \sum_{x\in\mathbf{F}:\; Tr(1/x)=0} e(ax) = -\,2 \sum_{x\in\mathbf{F}:\; Tr(1/x)=1} e(ax).$$

Proof. We first have

$$K(a) = \sum_{x,\; Tr(1/x)=0} e(ax) - \sum_{x,\; Tr(1/x)=1} e(ax).$$

Since

$$\sum_{x\in\mathbf{F}} e(ax) = 0 = \sum_{x,\ Tr(1/x)=0} e(ax) + \sum_{x,\ Tr(1/x)=1} e(ax),$$

we obtain the equality of the lemma using

$$\sum_{x,\ Tr(1/x)=0} e(ax) = -\sum_{x,\ Tr(1/x)=1} e(ax). \qquad \square$$

We also need the following sum $G'(a,b)$, which we introduced in [4], and which we call an *inverse cubic*:

$$G'(a,b) \ = \ \sum_{x\in\mathbf{F}^*} e\left(a\,x^3 \ + \ \frac{b}{x}\right), \quad a\in\mathbf{F}^*,\ b\in\mathbf{F}.$$

Here we also extend this sum to $x=0$, setting $bx^{-1}=0$ at the point $x=0$. Thus

$$G(a,b) \ = \ \sum_{x\in\mathbf{F}} e\left(a\,x^3 \ + \ \frac{b}{x}\right) \ = \ G'(a,b) \ + \ 1.$$

It is easy to check that $G(a,a) = G(a^2,a^2)$. This follows immediately from the equality $G'(a,a) = G'(a^2,a^2)$, which was given in [4]. We also have to bound these sums.

LEMMA 10. *Let m be even. For any $a,b\in\mathbf{F}$, where $(a,b)\neq(0,0)$, we have*

$$|G(a,b)| \le 2^{m/2+2}. \tag{2.5}$$

Proof. We gave an upper bound on $|G'(a,b)|$ in [4, Lemma 14] for odd m, but it is easy to check that our proof in [4] holds for even m too. This upper bound is as follows:

$$|G'(a,b)| \ \le \ 4\sqrt{2^m}.$$

Since $G(a,b)$ is a multiple of 4 for any $a,b\in\mathbf{F}$ and $G(a,b)=G'(a,b)+1$, the proof is completed. $\square$

Now, by the two next lemmas, we introduce some important relations linking partial sums with other sums considered here. To see the difference between even and odd cases, we formulate these results for both m, even and odd, and prove only the even cases. The odd cases are, respectively, Lemmas 10 and 12 in [4]. We mention that the partial sum $P(a,b)$, defined in [4], is not doubled (as here).

LEMMA 11. *Let a be any element of $\mathbf{F}^*$, where $\mathbf{F}$ has the order 2^m. Then*

$$P(a,a) \ = \ \begin{cases} K(a) \ + \ 2\,C(a,a) & if\ m\ is\ even, \\ K(a) & if\ m\ is\ odd. \end{cases}$$

Proof. Let m be even. We first have

$$\sum_{x\in\mathbf{F}} e(a(x^3+x)) = \sum_{x\in\mathbf{F},\ Tr(1/x)=0} e(a(x^3+x)) + \sum_{x\in\mathbf{F},\ Tr(1/x)=1} e(a(x^3+x))$$

which means

$$C(a,a) \ = \ \frac{1}{2}\,P(a,a) + \sum_{x\in\mathbf{F},\ Tr(1/x)=1} e(a(x^3+x)). \tag{2.6}$$

Moreover, in the case where m is even,

$$\{\ x^3+x \mid x\in\mathbf{F},\ Tr(1/x)=1\ \} = \{\ y\in\mathbf{F} \mid Tr(1/y)=1\ \}. \tag{2.7}$$

This is because

$$Tr\left(\frac{1}{x^3+x}\right) = Tr\left(\frac{1}{x}+\frac{1}{x+1}+\frac{1}{x^2+1}\right),$$

and the equation $x^3+x+c=0$ has a unique solution if and only if $Tr(1/c)=1$. We know that there are $M_1=2^{m-1}$ such c and then 2^{m-1} elements x^3+x (in the set above on the right) since for every such c

$$Tr\left(\frac{1}{c}\right) = Tr\left(\frac{1}{x^3+x}\right) = Tr\left(\frac{1}{x}\right) = 1$$

(see Lemmas 3 and 4). So, both sets in (2.7) have the same cardinality M_1. We deduce

$$\sum_{x\in\mathbf{F},\ Tr(1/x)=1} e(a(x^3+x)) = \sum_{y\in\mathbf{F},\ Tr(1/y)=1} e(ay).$$

Using (2.6) and Lemma 9, we get

$$P(a,a) \ = \ 2\,C(a,a) \ - \ 2\sum_{y\in\mathbf{F},\ Tr(1/y)=1} e(ay) \ = \ 2\,C(a,a) \ + \ K(a). \qquad \square$$

LEMMA 12. *For any $a\in\mathbf{F}^*$,*

$$P(a,0) \ = \ \begin{cases} G(a,a) \ + \ C(a) & if\ m\ is\ even,\\ G(a,a) & if\ m\ is\ odd. \end{cases}$$

Proof. Recall that we denote $C(a)=C(a,0)$. Also

$$P(a,0) = 2\sum_{x,\ Tr(1/x)=0} e(ax^3).$$

We have, using Lemma 5,

$$\sum_{x\in\mathbf{F}} e(ax^{-3}) + \sum_{x\in\mathbf{F}} e(ax^{-3}+x) = 2\sum_{x,\ Tr(x)=0} e(ax^{-3}) = P(a,0),$$

with

$$\sum_{x\in\mathbf{F}} e(ax^{-3}) = \sum_{x\in\mathbf{F}} e(ax^3) = C(a)$$

and, moreover,

$$\begin{aligned} G(a,1) &= \sum_{x\in\mathbf{F}} e(ax^3+x^{-1}) = \sum_{y\in\mathbf{F}} e(a^4y^3+y^{-1})\\ &= \sum_{z\in\mathbf{F}} e(az^3+az^{-1}) \ = \ G(a,a), \end{aligned}$$

with $y=x^4$ and $z=ay$. $\square$

3. Cosets of weight four in terms of nonlinear systems of equations. Let D be a coset of $BCH(2^m, 8)$ with syndrome $S = (a, b, c)$. To find the number of coset leaders in D, one needs to solve the following system of equations over $\mathbf{F}$:

$$\begin{aligned} x \;+\; y \;+\; z \;+\; u &= a, \\ x^3 \;+\; y^3 \;+\; z^3 \;+\; u^3 &= b, \\ x^5 \;+\; y^5 \;+\; z^5 \;+\; u^5 &= c. \end{aligned} \tag{3.1}$$

Here x, y, z, and u are pairwise distinct elements of $\mathbf{F}$. Here we are interested in cosets of weight 4 which are not contained in the Reed–Muller code of order $m-2$. That is, $\{x, y, z, u\}$ is not a 2-dimensional flat or, equivalently, $a \neq 0$ in (3.1). For the case of odd m, cosets which are contained in the Reed–Muller code of order $m-2$ have been described in [3]. The approach, which we used in [3] for odd m, can be used, of course, for the case of even m.

Denote by $\mu(a, b, c)$ the number of different solutions to the system (3.1), i.e., the number of unordered 4-sets of different elements x, y, z, u of $\mathbf{F}$, which satisfy (3.1). So, for fixed elements $a, b, c \in \mathbf{F}$, this number defines exactly the number of leaders of D.

We now recall some general properties of our system (3.1). They can be checked easily and have been considered for odd m in more detail but with another terminology in [3, Lemma 4.4].

PROPOSITION 1. *A 4-tuple $\{x, y, z, u\}$ is a solution to* (3.1) *for given (a, b, c) if and only if a 4-tuple $\{g\,x, g\,y, g\,z, g\,u\}$ is a solution to* (3.1) *for given (a', b', c'), where*

$$a' = g\,a, \qquad b' = g^3 b, \qquad c' = g^5 c, \qquad g \in \mathbf{F}^*.$$

PROPOSITION 2. *A 4-tuple $\{x, y, z, u\}$ is a solution to* (3.1) *for given (a, b, c) if and only if a 4-tuple $\{x+h, y+h, z+h, u+h\}$, $h \in \mathbf{F}$, is a solution to* (3.1) *for given (a', b', c'), where*

$$a' = a, \qquad b' = b + ha(h + a), \qquad c' = c + ha(h^3 + a^3).$$

PROPOSITION 3. *A 4-tuple $\{x, y, z, u\}$ is a solution to* (3.1) *for given (a, b, c) if and only if a 4-tuple $\{x^2, y^2, z^2, u^2\}$ is a solution to* (3.1) *for given (a', b', c'), where*

$$a' = a^2, \qquad b' = b^2, \qquad c' = c^2.$$

For fixed a, b, and c, denote by $V(a, b, c)$ the set of all 4-sets $\{x, y, z, u\}$ which are solutions to (3.1), i.e., in our notation $\#V(a, b, c) = \mu(a, b, c)$. Denote by $\mathcal{V}$ all of the sets of 4-sets, which are solutions to (3.1) for some a, b, c,

$$\mathcal{V} = \bigcup_{a \in \mathbf{F}^*,\ b, c \in \mathbf{F}} V(a, b, c).$$

This set $\mathcal{V}$ can be partitioned into different orbits, which are induced by applying Propositions 1–3.

DEFINITION 1. *For given elements $a, b, c \in \mathbf{F}$ we define the orbit $\mathcal{O}(a, b, c)$ as the set of $V(a', b', c')$, which can be obtained from $V(a, b, c)$ by all possible transformations given in Propositions* 1–3.

According to Propositions 1–3, all sets $V(a', b', c')$ from the orbit $\mathcal{O}(a, b, c)$ have the same cardinality $\mu(a, b, c)$. For arbitrary element $\eta \in \mathbf{F}$, we denote by $\ell_{\eta, m}$ the size

of the cyclotomic coset $C_\eta = \{\eta, \eta^2, \eta^{2^2}, \ldots\}$ of η induced by the action of Frobenius automorphisms of $\mathbf{F} = GF(2^m)$, i.e.,

$$\ell_{\eta,m} = \#C_\eta = \min\{s \mid s > 0,\ \eta^{2^s} = \eta\}.$$

Now we are going to prove that all orbits $\mathcal{O}(a,b,c)$ have a cardinality which depends on the value of $\ell_{\eta,m}$ only, for some η which is defined by the next lemma.

LEMMA 13. *Let a, b, c be arbitrary elements of $\mathbf{F}$, where $a \neq 0$. Let $\mu(a,b,c)$ be the number of solutions to the system* (3.1).

(i) *If $Tr(b/a^3) = 0$, then*

$$\mu(a,b,c) = \mu(1,0,\eta),$$

where

$$\eta = \frac{c}{a^5} + \frac{b^2}{a^6} + \frac{b}{a^3}. \tag{3.2}$$

(ii) *If $Tr(b/a^3) = 1$, then*

$$\mu(a,b,c) = \mu(1,\delta,\eta),$$

where δ is an arbitrary element of $\mathbf{F}^*$ with $Tr(\delta) = 1$ and where

$$\eta = \frac{c}{a^5} + \frac{b^2}{a^6} + \frac{b}{a^3} + \delta^2 + \delta.$$

Proof. Consider an arbitrary set $V(a,b,c)$, where a, b, c are arbitrary elements of $\mathbf{F}$ and $a \neq 0$. Using Proposition 1 with $g = 1/a$, we obtain the set $V(1, b/a^3, c/a^5)$, which has the same cardinality as $V(a,b,c)$. Now we apply Proposition 2 to this set. We obtain for any h,

$$V(1,\ b/a^3 + h(h+1),\ c/a^5 + h(h^3+1)).$$

First, assume that $Tr(b/a^3) = 0$. Consider the following quadratic equation on h:

$$h^2 + h + \frac{b}{a^3} \ = \ 0. \tag{3.3}$$

Since $Tr(b/a^3) = 0$, this equation has two distinct roots h_1 and h_2 in the field $\mathbf{F}$, and we choose any one of these roots as h. In such a way we obtain the set $V(1,0,\eta)$ where

$$\eta \ = \ \frac{c}{a^5} \ + \ h^4 \ + \ h. \tag{3.4}$$

Summing expression (3.3) and the expression obtained by squaring of (3.3), we arrive at the following formula for $h^4 + h$:

$$h^4 \ + \ h \ = \ \frac{b}{a^3} \ + \ \frac{b^2}{a^6},$$

which does not depend on the choice of the roots h_1 and h_2. Using this equality in (3.4), we obtain the formula (3.2) for η, given in Lemma 13 for the case (i).

Now consider the case (ii), when $Tr(b/a^3) = 1$. In this case (3.3) has no solutions in $\mathbf{F}$. Hence we cannot eliminate the element b/a^3, or even reduce it to 1. In this

case we cannot do anything better than choose $\delta \in F^*$ such that $Tr(\delta) = 1$, with h satisfying

$$h^2 + h + \frac{b}{a^3} + \delta = 0.$$

For any such element δ the equation above has two solutions, say h_1 and h_2. Hence, for a given b/a^3, we can take any element δ with $Tr(\delta) = 1$. Then we get the set $V(1, \delta, \eta)$, which has the same cardinality as $V(a, b, c)$. The expression for η is obtained in the same way as for η above. □

Note that for any i the set $V(1, 0, \eta^{2^i})$ belongs to the orbit $\mathcal{O}(1, 0, \eta)$, by definition of the orbits. Also, we have

$$V(1, \delta^{2^i}, \eta^{2^i}) \in \mathcal{O}(1, \delta, \eta).$$

Thus, according to Lemma 13, the set $\mathcal{V}$ is partitioned into the orbits of two types: $\mathcal{O}(1, 0, \eta)$ and $\mathcal{O}(1, \delta, \eta)$. We are going to compute the cardinality of these orbits. Our next proposition, together with Lemma 13, gives the length of any orbit $\mathcal{O}(a, b, c)$.

PROPOSITION 4. *The parameters η and δ are defined by Lemma* 13. *The length of the orbit $\mathcal{O}(1, 0, \eta)$ and the length of the orbit $\mathcal{O}(1, \delta, \eta)$ only depend on the size $\ell_{\eta,m}$ of the cyclotomic coset C_η of η. More precisely,*

$$\#\mathcal{O}(1, 0, \eta) = \#\mathcal{O}(1, \delta, \eta) = (2^m - 1)2^{m-1}\ell_{\eta,m}.$$

Proof. First, note that by Lemma 13 we proved that for any (a, b, c), the set $V(a, b, c)$ is either in $\mathcal{O}(1, 0, \eta)$ or $\mathcal{O}(1, \delta, \eta)$, for some δ such that $Tr(\delta) = 1$, where η is uniquely defined.

Let η be any element of $\mathbf{F}$. According to Definition 1, we have to count the number of distinct sets $V(a, b, c)$ which belong to $\mathcal{O}(1, 0, \eta)$. We can choose in 2^{m-1} ways an element $\beta \in \mathbf{F}$ and, further, the element $a \in \mathbf{F}^*$ in $2^m - 1$ ways. To be clear, we proceed as follows:

$$(1, 0, \eta) \longrightarrow (1, \beta = h^2 + h, \eta + h^4 + h) \longrightarrow (a, \beta a^3, (\eta + \beta + \beta^2)a^5)$$

and obtain $(2^m - 1)2^{m-1}$ different triples

$$(a, b, c), \quad b = \beta a^3, \quad \text{and} \quad c = (\eta + \beta + \beta^2)a^5.$$

Moreover, for each such triple, the sets $V(a, b, c_i)$ with $c_i = (\eta^{2^i} + \beta + \beta^2)a^5$ also belong to $\mathcal{O}(1, 0, \eta)$, which allow us to get at all $(2^m - 1)2^{m-1}\ell_{\eta,m}$ elements.

We proceed in the same way to count the number of distinct sets $V(a, b, c)$ which belong to $\mathcal{O}(1, \delta, \eta)$ (where $Tr(\delta) = 1$). We have, as before

$$(1, \delta, \eta) \rightarrow (1, \beta = \delta + h^2 + h, \eta + h^4 + h) \rightarrow (a, \beta a^3, (\eta + \beta + \delta + (\beta + \delta)^2)a^5)$$

and then $(2^m - 1)2^{m-1}$ different triples

$$(a, b, c), \quad b = \beta a^3, \quad \text{and} \quad c = (\eta + \beta + \delta + (\beta + \delta)^2)a^5.$$

Note that the image of the map $h \mapsto h + h^2 + \delta$ is the set of all β such that $Tr(\beta) = 1$. This image does not depend on δ. We have to take into account that $V(1, \delta^{2^i}, \eta^{2^i})$ belongs to $\mathcal{O}(1, \delta, \eta)$ for any i. Due to our previous remark, we have to consider only the length of C_η, providing that the cardinality of $\mathcal{O}(1, \delta, \eta)$ equals $(2^m - 1)2^{m-1}\ell_{\eta,m}$. □

Remark 1. In this section, we assume that $a \neq 0$ for the study of $\mu(a,b,c)$. When $a = 0$ then the corresponding coset, say D, is contained in the Reed–Muller code of order $m-2$. According to Proposition 3, it is clear that *if* $\{x,y,z,u\}$ *is a solution to* (3.1) *for given* $(0,b,c)$*, then any* 4*-tuple* $\{x+h, y+h, z+h, u+h\}$ *is a solution too, for any* $h \in \mathbf{F}$. In this case the coset D is such that each coordinate position is covered by at least one leader of D. Since the weight of D is 4, the supports of two leaders cannot intersect, proving that the number of leaders is 2^{m-2}. Since any leader of D is a minimum codeword of the Reed–Muller code of order $m-2$, its support is an affine subspace of dimension 2. As there are $(2^m-1)(2^m-2)/6$ linear subspaces of dimension 2, there are also the same number of cosets of B of weight 4 corresponding to triples of the form $(0,b,c)$.

4. On the number of solutions to the system of equations and exponential sums. The main result of this paper is the following explicit expression for the number of solutions to the system (3.1) in terms of four different types of exponential sums. We repeat the corresponding result from [4] for odd m and a new result for even m as one theorem (for completeness and to see the difference between these two cases).

THEOREM 1. *Let* $\mu(a,b,c)$ *be the number of different* 4*-sets* $\{x,y,z,u\}$*, where* x,y,z,u *are pairwise distinct elements of* $\mathbf{F}$*, which are solutions to the system* (3.1)*, where* $a,b,$ *and* c *are arbitrary elements of a field* $\mathbf{F}$ *of cardinality* 2^m $(m \geq 4)$ *and* $a \neq 0$*. Let*

$$\epsilon = Tr\left(\frac{b}{a^3}\right) \quad \textit{and} \quad \lambda = \frac{b}{a^3}\left(\frac{b}{a^3}+1\right) + \frac{c}{a^5} + 1. \tag{4.1}$$

If $\lambda \neq 0$*, then*

$$\mu(a,b,c) \;=\; \mu(\epsilon,\lambda) \;=\; \frac{1}{3}\,M(\epsilon,\lambda) \tag{4.2}$$

where $M(\epsilon,\lambda)$ *is even and equal to: for even* m

$$\begin{aligned} 8\,M(\epsilon,\lambda) &= 2^m \;-\; 8 \;+\; 3G(\lambda,\lambda) \;+\; C(\lambda) \\ &\quad + (-1)^{\epsilon}\,(2K(\lambda) \;+\; 4C(\lambda,\lambda) \;-\; 8)\,, \end{aligned} \tag{4.3}$$

and for odd m

$$\begin{aligned} 8\,M(\epsilon,\lambda) &= 2^m \;-\; 8 \;+\; 3G(\lambda,\lambda) \\ &\quad + (-1)^{\epsilon+1}\,(2K(\lambda) \;+\; 2C(\lambda,\lambda) \;-\; 8)\,. \end{aligned} \tag{4.4}$$

If $\lambda = 0$*, then*

$$\mu(\epsilon,0) \;=\; 0.$$

We want to solve the system (3.1) for the general case $Tr(b/a^3) = \epsilon$. Thus, we do not use the reduced form $\mathcal{O}(1,0,\eta)$ or $\mathcal{O}(1,\delta,\eta)$ of the orbits of solutions $\mathcal{O}(a,b,c)$, obtained in the previous section. For our purposes we consider the system (3.1) in the following form:

$$\begin{aligned} x \;+\; y \;+\; z \;+\; u &= 1, \\ x^3 \;+\; y^3 \;+\; z^3 \;+\; u^3 &= b', \\ x^5 \;+\; y^5 \;+\; z^5 \;+\; u^5 &= c', \end{aligned} \tag{4.5}$$

where x, y, z, and u are pairwise distinct elements of $\mathbf{F}$ and where $b' = b/a^3$ and $c' = c/a^5$ are arbitrary elements of $\mathbf{F}$. From now on, we use the following notation:

$$\mathbf{F}^{**} = \mathbf{F} \setminus \{0, 1\}.$$

Before we begin to prove the theorem we give one simple lemma and several statements, which reduce some exponential sums to the sums, which we introduced in section 2.

LEMMA 14. *Let* $\{x, y, z, u\}$ *be a solution to* (4.5). *Then a* 4*-set* $\{x+1, y+1, z+1, u+1\}$ *is a solution to* (4.5).

Proof. The proof follows by direct checking. ∎

Define three following functions $g_i(v)$ from $\mathbf{F}^{**}$ to $\mathbf{F}^{**}$:

$$\begin{aligned} g_1(v) &= \lambda\left(\frac{v+1}{v^3}\right), \\ g_2(v) &= \lambda\left(\frac{v}{(v+1)^3}\right), \\ g_3(v) &= \lambda\left(\frac{1}{v} + \frac{1}{v+1}\right). \end{aligned}$$

Denote by $S(g)$ the following exponential sum:

$$S(g) = \sum_{v \in \mathbf{F}^{**}} e(g(v)).$$

PROPOSITION 5. *Let* $\lambda \neq 0$. *Then*

$$S(g_1) = S(g_2) = C(\lambda, \lambda) - 2.$$

Proof. Since $g_1(v) = g_2(v+1)$, we have that $S(g_1) = S(g_2)$. Consider $S(g_1)$:

$$\begin{aligned} S(g_1) &= \sum_{v \in \mathbf{F}^{**}} e\left(\lambda \frac{v+1}{v^3}\right) \\ &= \sum_{v \in \mathbf{F}^{**}} e\left(\frac{\lambda}{v^2} + \frac{\lambda}{v^3}\right) \\ &= \sum_{\xi \in \mathbf{F}^{**}} e(\lambda(\xi^3 + \xi^2)) \\ &= \sum_{\zeta \in \mathbf{F}^{**}} e(\lambda(\zeta^3 + \zeta)) \\ &= C(\lambda, \lambda) - 2, \end{aligned}$$

where we twice changed the variable $v = 1/\xi$ and $\xi = \zeta + 1$. ∎

PROPOSITION 6. *Let* $\lambda \neq 0$. *Then*

$$S(g_3) = K(\lambda) - 2.$$

Proof. This result is an instance of [7, Theorem 1]. We briefly give the proof for

clarity and completeness:

$$\begin{aligned} S(g_3) &= \sum_{v\in\mathbf{F}^{**}} e\left(\frac{\lambda}{v}+\frac{\lambda}{v+1}\right) \\ &= \sum_{h\in\mathbf{F}^{**}} e\left(\frac{\lambda h^2}{h+1}\right) \\ &= \sum_{h\in\mathbf{F}^{**}} e\left(\lambda\,(h+1)+\frac{\lambda}{h+1}\right) \\ &= K(\lambda) \;-\; 2, \end{aligned}$$

where $h = 1/v$ and using (2.3). ∎

PROPOSITION 7. *Let $\lambda \neq 0$. Then*

$$\begin{aligned} S(g_1+g_2+g_3) &= P(\lambda,\lambda) \;-\; 2 \\ &= 2C(\lambda,\lambda) \;+\; K(\lambda) \;-\; 2. \end{aligned}$$

Proof. The partial sum P is defined by (2.1). First, we reduce $S(g_1+g_2+g_3)$ to the simplified form as follows:

$$\begin{aligned} g_1+g_2+g_3 &= \lambda\left(\frac{v+1}{v^3}+\frac{v}{(v+1)^3}+\frac{1}{v}+\frac{1}{v+1}\right) \\ &= \lambda\left(\frac{1}{(v^2+v)^3}+\frac{1}{v^2+v}\right). \end{aligned}$$

Changing the variable $v^2+v=\xi$ with $Tr(\xi)=0$, we obtain

$$\begin{aligned} S(g_1+g_2+g_3) &= \sum_{v\in\mathbf{F}^{**}} e\left(\lambda\left(\frac{1}{(v^2+v)^3}+\frac{1}{v^2+v}\right)\right) \\ &= 2\sum_{\xi\in\mathbf{F}^{*}:\ Tr(\xi)=0} e\left(\lambda\left(\frac{1}{\xi^3}\;+\;\frac{1}{\xi}\right)\right) \\ &= 2\sum_{\zeta\in\mathbf{F}^{*}:\ Tr(1/\zeta)=0} e\left(\lambda(\zeta^3+\zeta)\right) \\ &= P(\lambda,\lambda) \;-\; 2. \end{aligned}$$

Here we have to explain why we return to summing over $\mathbf{F}^*$, but not $\mathbf{F}^{**}$ as we started. Indeed, $Tr(1)=0$, hence the equation $v^2+v=1$ always has a solution in $\mathbf{F}$, the field of order 2^m, for even m. Therefore, when we change v^2+v $(v\in\mathbf{F}^{**})$ to ξ we have to extend $\mathbf{F}^{**}$ into $\mathbf{F}^*$. Now using Lemma 11 we obtain the final expression. ∎

PROPOSITION 8. *Let $\lambda \neq 0$. Then*

$$\begin{aligned} S(g_1+g_2) &= P(\lambda,0) \;-\; 2 \\ &= C(\lambda) \;+\; G(\lambda,\lambda) \;-\; 2. \end{aligned}$$

Proof. We have

$$\begin{aligned} g_1+g_2 &= \lambda\left(\frac{v+1}{v^3}\;+\;\frac{v}{(v+1)^3}\right) \\ &= \lambda\left(\frac{1}{(v^2+v)^3}\right) \\ &= \frac{\lambda}{\xi^3} \;=\; \lambda\,\zeta^3, \end{aligned}$$

where we change variables $v^2 + v = \xi$ and then $1/\xi = \zeta$. Taking into account that $Tr(v^2 + v) = Tr(\xi) = Tr(1/\zeta) = 0$, we rewrite $S(g_1 + g_2)$ as follows:

$$\begin{aligned} S(g_1 + g_2) &= 2 \sum_{\zeta \in \mathbf{F}^*:\ Tr(1/\zeta)=0} e\,(\lambda \zeta^3) \\ &= P(\lambda, 0) \ - \ 2. \end{aligned}$$

Then we obtain the final expression using Lemma 12. □

PROPOSITION 9. *Let* $\lambda \neq 0$. *Then*

$$S(g_1 + g_3) \ = \ S(g_2 + g_3) \ = \ G(\lambda, \lambda) \ - \ 2.$$

Proof. Since $g_1(v) = g_2(v + 1)$ and $g_3(v) = g_3(v + 1)$ we deduce that $S(g_1 + g_3) = S(g_2 + g_3)$. So it is enough to compute $S(g_1 + g_3)$. First, we rewrite $g_1 + g_3$ as follows:

$$\begin{aligned} g_1 \ + \ g_3 &= \lambda \left(\frac{v+1}{v^3} \ + \ \frac{1}{v} + \frac{1}{v+1} \right) \\ &= \lambda \left(1 + \frac{v^2 + v + 1}{v^3} \ + \ 1 \ + \ \frac{1}{v+1} \right) \\ &= \lambda \left(\frac{(v+1)^3}{v^3} \ + \ \frac{v}{v+1} \right). \end{aligned}$$

Obviously, the mapping $v \mapsto (v + 1)/v$ is a 1-to-1 mapping from $\mathbf{F}^{**}$ onto $\mathbf{F}^{**}$. Therefore, changing $\xi = (v + 1)/v$, we obtain for $S(g_1 + g_3)$:

$$\begin{aligned} S(g_1 + g_3) &= \sum_{\xi \in \mathbf{F}^{**}} e \left(\lambda \left(\xi^3 \ + \ \frac{1}{\xi} \right) \right) \\ &= G(\lambda, \lambda) \ - \ 2. \end{aligned}$$ □

The proof of Theorem 1. Solving the system (4.5), we will, for short, use b and c during the proof instead of b' and c'.

We introduce two new variables

$$x \ + \ y \ = \ v, \quad xy \ = \ w.$$

As x, y, z, and u are all different, the element v belongs to the set $\mathbf{F}^{**}$. Using these new variables we can express $x^3 + y^3$ as follows:

$$x^3 \ + \ y^3 \ = \ v^3 \ + \ wv. \tag{4.6}$$

As $z + u = v + 1$ and $z^3 + u^3 = (v + 1)^3 + z\,u(v + 1)$ we can obtain from the second line of (4.5) that

$$w\,v \ + \ zu(v+1) \ = \ v^2 \ + \ v \ + \ b \ + \ 1. \tag{4.7}$$

Now we want, using the third line of (4.5), to obtain an expression similar to (4.7), which includes only new variables v and w and also the product $z\,u$. We have from (4.6) and the second line of (4.5)

$$\begin{aligned} (x+y)^5 &= x^5 \ + \ y^5 \ + \ xy(x^3 + y^3) \\ &= x^5 \ + \ y^5 \ + \ w(v^3 + w\,v) = v^5 \end{aligned}$$

and

$$\begin{aligned}(z+u)^5 &= z^5+u^5 \ + \ zu(z^3+u^3)\\ &= z^5 \ + \ u^5 \ + \ zu(x^3+y^3+b)\\ &= z^5 \ + \ u^5 \ + \ zu(v^3+wv+b) = (v+1)^5.\end{aligned}$$

Using these two expressions above and the third line of (4.5), we obtain

$$w^2\,v \ + \ w\,v^3 \ + \ zu\,(v^3+wv+b) \ = \ v^4 \ + \ v \ + \ c \ + \ 1. \tag{4.8}$$

We multiply (4.8) by $v+1$ and replace $z\,u$ by its value in (4.7), Thus, we get the following quadratic equation for w:

$$w^2v \ + \ w\,(v^2+v) \ + \ (v+1)(v^4+v+c+1) \ + \ (v^2+v+b+1)(v^3+b) \ = \ 0,$$

which gives, with $\lambda = c+1+b(b+1)$,

$$w^2 \ + \ w\,(v+1) \ + \ (v^2+v)(b+1)+b+c \ + \ \frac{\lambda}{v} \ = \ 0. \tag{4.9}$$

As we know from Lemma 2, this equation has two different roots in $\mathbf{F}$ if and only if

$$Tr\left(\frac{(v^2+v)(b+1)+b+c \ + \ \lambda/v}{(v+1)^2}\right) \ = \ 0. \tag{4.10}$$

Denote by $w_1 = w_1(v)$ and $w_2 = w_2(v)$ two distinct roots of (4.9). Now we return to the beginning of our proof. Two equalities $x \ + \ y \ = \ v$ and $x\,y \ = \ w_i$, $i \in \{1,2\}$, as well as two equalities $z+u = v+1$ and $z\,u = (w_i\,v+v^2+v+b+1)/(v+1)$ imply the two following trace conditions (Lemma 1):

$$Tr\left(\frac{w_i}{v^2}\right) \ = \ 0 \tag{4.11}$$

and

$$Tr\left(\frac{w_i\,v+v(v+1)+b+1}{(v+1)^3}\right) \ = \ Tr\left(\frac{w_i\,v+b+1}{(v+1)^3}\right) \ = \ 0. \tag{4.12}$$

As $w_1+w_2 = v+1$, it is easy to see that the validity of both conditions of (4.11) for one of w_i implies the validity of these conditions for the other.

Recall Lemma 14. Assume that (x,y,z,u) is a solution to (4.5) corresponding to $w_1 = w_1(v)$. Then it is easy to see that a 4-tuple $(x+1, y+1, z+1, u+1)$ is a solution to (4.5) corresponding to $w_2 = w_2(v)$.

Now we want to rewrite the conditions (4.11) and (4.12) in a more acceptable form. More exactly, using the fact that w_1 and w_2 are the roots of the quadratic equation (4.9), we want to eliminate w_i from the conditions (4.11) and (4.12). We

start from the first condition (let $w_i = w$):

$$\begin{aligned} Tr\left(\frac{w}{v^2}\right) &= Tr\left(\frac{w + wv + wv}{v^2}\right) = Tr\left(\frac{w(v+1)}{v^2} \ + \ \frac{w}{v}\right) \\ &= Tr\left(\frac{w(v+1) + w^2}{v^2}\right) \\ &= Tr\left(\frac{(v^2+v)(b+1) + b + c \ + \ \lambda/v}{v^2}\right) \\ &= Tr\left(b \ + \ \frac{b+1}{v} + \frac{b+c}{v^2} + \frac{\lambda}{v^3}\right) \\ &= Tr\left(b \ + \ \frac{(c+1) + b(b+1)}{v^2} + \frac{\lambda}{v^3}\right) \\ &= Tr\left(b \ + \ \lambda\left(\frac{v+1}{v^3}\right)\right), \end{aligned}$$

where we used the condition (4.9), that $Tr(x) = Tr(x^2)$ and $Tr(1) = 0$, since m is even. Thus (4.11) can be written as follows:

$$Tr\left(\lambda\left(\frac{v+1}{v^3}\right) \ + \ b\right) \ = \ 0. \tag{4.13}$$

Now we have for the condition (4.12):

$$\begin{aligned} Tr\left(\frac{wv + b + 1}{(v+1)^3}\right) &= Tr\left(\frac{wv + b + 1 + w + w}{(v+1)^3}\right) \\ &= Tr\left(\frac{w + b + 1}{(v+1)^3} + \frac{w}{(v+1)^2}\right) \\ &= Tr\left(\frac{w^2 + w(v+1) + (v+1)(b+1)}{(v+1)^4}\right) \\ &= Tr\left(\frac{(v+1)(b+1) + (v^2+v)(b+1) + b + c + \lambda/v}{(v+1)^4}\right) \\ &= Tr\left(\frac{\lambda}{v(v^4+1)} + \frac{(b+1)(v^2+1) + b + c}{v^4+1}\right) \\ &= Tr\left(\frac{\lambda}{v(v^4+1)} + \frac{b+1}{v^2+1} + \frac{b+c}{v^4+1}\right) \\ &= Tr\left(\frac{\lambda}{v(v^4+1)} + \frac{\lambda}{v^4+1}\right) = Tr\left(\frac{\lambda}{v(v+1)^3}\right). \end{aligned}$$

But

$$\frac{1}{v(v+1)^3} = \frac{1}{v} + \frac{1}{v+1} + \frac{1}{(v+1)^2} + \frac{1}{(v+1)^3}.$$

Hence we can rewrite the condition (4.12) as follows:

$$Tr\left(\lambda\left(\frac{v}{(v+1)^3} \ + \ \frac{1}{v} \ + \ \frac{1}{v+1}\right)\right) \ = \ 0. \tag{4.14}$$

Now we rewrite the condition (4.10). We have

$$\begin{aligned} & Tr\left(\frac{(v^2+v)(b+1)+b+c\ +\ \lambda/v}{(v+1)^2}\right) \\ &= Tr\left(b+\frac{b+1}{v+1}+\frac{b+c}{(v+1)^2}+\frac{\lambda}{v(v+1)^2}\right) \\ &= Tr\left(b+\frac{\lambda}{(v+1)^2}+\frac{\lambda}{v(v+1)^2}\right), \end{aligned}$$

using $v^2+v=v^2+1+v+1$ and properties of the trace function.

Thus the condition (4.10) is equivalent to the following condition:

$$Tr\left(b\ +\ \lambda\left(\frac{1}{v}\ +\ \frac{1}{v+1}\right)\right)=0. \tag{4.15}$$

We continue the proof of the theorem. So, in order to find the number $\mu(1,b,c)=\mu(\epsilon,\lambda)$ we have to find the following number, which we denote by $M(\epsilon,\lambda)$: *how many times all three conditions* (4.13), (4.14), *and* (4.15) *are simultaneously satisfied when* v *runs over* $\mathbf{F}^{**}$. It is easy to write the expression for the number $M(\epsilon,\lambda)$ in terms of exponential sums. Denote that (recall that $\lambda=\eta+1=b(b+1)+c+1$)

$$\begin{aligned} f_1 &= \lambda\left(\frac{v+1}{v^3}\right)\ +\ b, \\ f_2 &= \lambda\left(\frac{v}{(v+1)^3}\ +\ \frac{1}{v}\ +\ \frac{1}{v+1}\right), \\ f_3 &= \lambda\left(\frac{1}{v}\ +\ \frac{1}{v+1}\right)\ +\ b. \end{aligned}$$

By the definition we have

$$M(\epsilon,\lambda)=\frac{1}{8}\sum_{v\in\mathbf{F}^{**}}\left(1+(-1)^{Tr(f_1)}\right)\left(1+(-1)^{Tr(f_2)}\right)\left(1+(-1)^{Tr(f_3)}\right).$$

Multiplying into the parentheses and using our notation $e(a)=(-1)^{Tr(a)}$ and $\epsilon=Tr(b)$, we obtain

$$\begin{aligned} 8\,M(\epsilon,\lambda) = & \sum_{v\in\mathbf{F}^{**}}1+\sum_{v\in\mathbf{F}^{**}}e(f_1)+\sum_{v\in\mathbf{F}^{**}}e(f_2) \\ & +\sum_{v\in\mathbf{F}^{**}}e(f_3)+\sum_{v\in\mathbf{F}^{**}}e(f_1+f_2)+\sum_{v\in\mathbf{F}^{**}}e(f_1+f_3) \\ & +\sum_{v\in\mathbf{F}^{**}}e(f_2+f_3)+\sum_{v\in\mathbf{F}^{**}}e(f_1+f_2+f_3). \end{aligned}$$

Recall that $S(g)$ denotes the following exponential sum of g:

$$S(g)\ =\ \sum_{v\in\mathbf{F}^{**}}e(g(v)).$$

Introducing the following notation:

$$\begin{aligned} S_i &= S(f_i),\quad i=1,2,3, \\ S_{i,j} &= S(f_i+f_j),\quad i\neq j,\ i,j\in\{1,2,3\}, \\ S_{1,2,3} &= S(f_1+f_2+f_3), \end{aligned}$$

we obtain for the number $M(\epsilon, \eta)$,

$$8\,M(\epsilon,\lambda) = 2^m - 2 + S_1 + S_2 + S_3 + S_{1,2} + S_{1,3} + S_{2,3} + S_{1,2,3}.$$

Using the three functions $g_i(v)$, $i = 1, 2, 3$, introduced previously, and the fact that $g_2(v) = g_1(v+1)$, our separate sums can be written as follows:

$$\begin{aligned} S_1 &= S(g_1(v) + b), \\ S_2 &= S(g_2(v) + g_3(v)), \\ S_3 &= S(g_3(v) + b), \\ S_{1,2} &= S(g_1(v) + g_2(v) + g_3(v) + b), \\ S_{1,3} &= S(g_1(v) + g_3(v)), \\ S_{2,3} &= S(g_2(v) + b), \\ S_{1,2,3} &= S(g_1(v) + g_2(v)). \end{aligned}$$

Since $S(g + b) = -S(g)$ for the case $\epsilon = Tr(b) = 1$ and $S(g + b) = S(g)$ for the case $\epsilon = Tr(b) = 0$, we arrive at the following expression for the number $M(\epsilon, \lambda)$:

$$\begin{aligned} 8\,M(\epsilon,\lambda) &= 2^m - 2 + S(g_1 + g_2) + S(g_1 + g_3) + S(g_2 + g_3) \\ &\quad + (-1)^{\epsilon}\left(S(g_1) + S(g_2) + S(g_3) + S(g_1 + g_2 + g_3)\right). \end{aligned} \tag{4.16}$$

Using Propositions 5–9 for all of the sums in (4.16), and recalling our initial notation

$$\lambda = b'(b'+1) + c' + 1 = \frac{b}{a^3}\left(\frac{b}{a^3} + 1\right) + \frac{c}{a^5} + 1 = \eta + 1,$$

we obtain the expression for $M(\epsilon, \lambda)$ in the theorem for the case of even m. It remains to prove (4.2). When we introduce the new variables $v = x + y$ and $w = xy$, we could choose x and y in 6 different ways from the four variables x, y, z, u. But it is easy to see that two "opposite" choices of the new variables: $v = x+y$, $w = xy$, and $v = z+u$, $w = zu$ result in the same quadratic equation (4.9) for w. Of course it is possible to say the same about choices $v = x + z$, $w = x\,z$ and $v = y + u$, $w = y\,u$ (respectively, $v = x + u$, $w = x\,u$, and $v = y + z$, $w = y\,z$).

This means that for each proper value of $v \in \mathbf{F}^{**}$ (when all three trace conditions (4.10), (4.13), and (4.14) are satisfied), we obtain a solution $\{x, y, z, u\}$ as well as a solution $\{z, u, x, y\}$ (note that here any solution $\{x, y, z, u\}$ we consider is up to permutations between x and y and between z and u). Therefore, when v runs over $\mathbf{F}^{**}$ each solution $\{x, y, z, u\}$ occurs exactly three times. In other words, three distinct proper values of v result in the same solutions, namely $\{x, y, z, u\}$, $\{x, z, y, u\}$, and $\{x, u, y, z\}$. This means that

$$\mu(\epsilon,\lambda) \;=\; \frac{1}{3}\,M(\epsilon,\lambda),$$

i.e., we obtain the equality (4.2). The integer $M(\epsilon, \lambda)$ is even according to Lemma 14.

Now consider the case where $\lambda = 0$ or $c + 1 = b^2 + b$ (we again use the short notation b and c instead of b' and c'). For this case the trace conditions (4.13), (4.14), and (4.15) reduce, respectively, to

$$Tr(b) = 0,\ Tr(0) = 0, \quad \text{and} \quad Tr(b) = 0.$$

TABLE 1
$m = 6;\ p(x) = x^6 + x + 1.$

λ	$Tr(\lambda)$	C	C_0	K	G	$\mu(0,\lambda)$	$\mu(1,\lambda)$
1	0	0	16	− 8	0	2	4
α	0	8	− 8	0	− 8	2	0
α^3	0	0	16	− 8	0	2	4
α^5	1	8	− 8	12	0	4	0
α^7	0	8	− 8	0	8	4	2
α^9	0	16	16	8	0	6	0
α^{11}	1	− 8	− 8	− 4	16	2	6
α^{13}	0	− 8	− 8	8	− 8	0	2
α^{15}	1	0	16	4	− 8	2	2
α^{21}	1	− 16	16	12	8	2	6
α^{23}	1	8	− 8	− 12	0	2	2
α^{27}	0	0	16	16	16	6	4
α^{31}	1	− 8	− 8	− 4	0	0	4

Therefore, for the case $\epsilon = Tr(b) = 1$ our system (4.5) has no solutions, i.e., $\mu(1,0) = 0$. We proceed now with the case $\epsilon = Tr(b) = 0$. According to Lemma 13 (with $a = 1$), we have

$$\mu(1,b,c) = \mu(1,0,c+b^2+b) = \mu(1,0,1),$$

since $b^2 + b = c + 1$. Now consider the system (4.5) with $b = 0$ and $c = 1$. It is easy to check that $\{\beta, \beta^2, 0, 0\}$, where $\beta \in \mathbf{F}_4$ is of order 3, is a solution of (4.5). We deduce that the coset of syndrome $(a,b,c) = (1,0,1)$ has minimum weight 2 and then cannot contain any codeword of weight 4, *i.e., there is no solution of* (4.5) *composed of* 4 *pairwise distinct elements of* $\mathbf{F}$. This completes the proof of Theorem 1. □

As a direct corollary of Theorem 1, we obtain the following lower and upper bounds for the number $\mu(a,b,c)$, *i.e., the number of coset leaders in any coset* D *of weight* 4 *with syndrome* (a,b,c) *with* $a \neq 0$. We use the bounds for the exponential sums $K(\lambda)$, $G(\lambda,\lambda)$ and $C(\lambda)$, $C(\lambda,\lambda)$, involved in the number of solutions $\mu(a,b,c)$ (see Lemma 6, (2.4), and (2.5)).

THEOREM 2. *Let* a,b,c $(a \neq 0)$ *be any elements of* $\mathbf{F}$ *where* $\mathbf{F}$ *is the finite field of order* 2^m*, with* m *even and* $m \geq 10$. *Let* λ *be defined as in* (4.1). *If* λ *is a cube, then*

$$2^m - 8 - 26\sqrt{2^m} \leq 24\,\mu(a,b,c) \leq 2^m + 26\sqrt{2^m}.$$

If further $T_2^m((\lambda)^{2/3}) \neq 0$*, then*

$$2^m - 8 - 18\sqrt{2^m} \leq 24\,\mu(a,b,c) \leq 2^m + 18\sqrt{2^m}.$$

If λ *is not a cube, then*

$$2^m - 8 - 21\sqrt{2^m} \leq 24\,\mu(a,b,c) \leq 2^m + 21\sqrt{2^m}.$$

We note that the second bound is better than the corresponding bounds for odd m, obtained in [4] and [8].

TABLE 2
$m = 8;\ p(x) = x^8 + x^7 + x^6 + x + 1.$

λ	$Tr(\lambda)$	C	C_0	K	G	$\mu(0,\lambda)$	$\mu(1,\lambda)$
1	0	− 32	− 32	32	32	10	16
α	0	16	16	− 16	− 16	10	8
α^3	0	0	− 32	− 8	32	12	14
α^5	1	16	16	20	8	16	8
α^7	0	16	16	8	− 16	12	6
α^9	1	0	− 32	28	− 8	10	6
α^{11}	1	16	16	− 28	8	12	12
α^{13}	0	− 16	16	0	0	8	14
α^{15}	1	0	− 32	4	− 24	6	6
α^{17}	0	16	16	32	− 16	14	4
α^{19}	0	16	16	8	16	16	10
α^{21}	1	− 32	− 32	− 4	8	4	16
α^{23}	0	− 16	16	− 24	16	8	18
α^{25}	0	16	16	− 16	16	14	12
α^{27}	0	0	− 32	16	− 16	8	6
α^{29}	1	16	16	− 4	− 8	12	8
α^{31}	0	− 16	16	0	0	8	14
α^{37}	0	− 16	16	24	0	10	12
α^{39}	1	32	− 32	− 12	− 8	12	4
α^{43}	1	16	16	20	− 8	14	6
α^{45}	0	0	− 32	16	16	12	10
α^{47}	1	− 16	16	12	8	10	14
α^{51}	0	32	− 32	24	16	18	4
α^{53}	1	16	16	− 4	24	16	12
α^{55}	1	16	16	− 4	8	14	10
α^{59}	0	− 16	16	− 24	0	6	16
α^{61}	1	− 16	16	− 12	− 40	2	10
α^{63}	1	0	− 32	4	8	10	10
α^{85}	0	16	16	− 16	48	18	16
α^{87}	1	0	− 32	− 20	− 8	6	10
α^{91}	1	− 16	16	− 12	8	8	16
α^{95}	1	− 16	16	12	− 8	8	12
α^{111}	0	0	− 32	− 8	0	8	10
α^{119}	0	16	16	8	32	18	12
α^{127}	0	− 16	16	24	− 16	8	10

5. Numerical results. We present in Tables 1 and 2 the values of all exponential sums involved in the expression of $\mu(a, b, c)$ for $m = 6$ and $m = 8$. In Tables 1 and 2, the results are given for a set of representatives of the cyclotomic cosets only (since it is the same for all elements from such coset). We distinguish for a given λ two cases: $\epsilon = 0$ or $\epsilon = 1$ (with notation of Theorem 1). So for each value λ we give two numbers $\mu(1, \lambda)$ and $\mu(0, \lambda)$. For short, we use the following notation: $K = K(\lambda)$, $C = C(\lambda, \lambda)$, $C_0 = C(\lambda, 0)$, and $G = G(\lambda, \lambda)$. We denote by $p(x)$ the primitive

polynomial generating $\mathbf{F}$.

6. Conclusion. In this paper, we extended to the even case our work [4] on the coset leaders of cosets of weight 4 of the codes $BCH(2^m, 8)$. By Theorem 1 we summarized our results for both cases, m even and m odd. Recall that we gave in [6] the coset distribution of all codes $BCH(2^m, 8)$.

Now, the main open problem remains the computation of the weight distribution of all cosets. It has be shown for odd m that all is known as soon as the numbers $\mu(\epsilon, \lambda)$, and the number of times they occur, are known [3]. We conjecture that this property holds for even m. We introduced lower and upper bounds for the number of coset leaders of cosets of weight 4. We conjectured in [3] that this number takes all values between its bounds, up to some divisibility property. This conjecture was disproved in [8]. So the first question is: Which values are suitable?

New properties of exponential sums K, G, and C arise from formula (4.3) and (4.4) and from elements of their proofs. We developed this aspect in the odd case [5]. In the even case, the relations between K and C are more interesting since the spectrum of C is more complicated. We will study this fact in a forthcoming paper.

REFERENCES

[1] E. R. Berlekamp, H. Rumsey, and G. Solomon, *On the solution of algebraic equations over finite fields*, Information and Control, 12 (1967) pp. 553–564.
[2] L. Carlitz, *Explicit evaluation of certain exponential sums*, Math. Scand., 44 (1979), pp. 5–16.
[3] P. Charpin and V. A. Zinoviev, *On coset weight distributions of the* 3*-error-correcting BCH-codes*, SIAM J. Discrete Math., 10 (1997), pp. 128–145.
[4] P. Charpin, T. Helleseth, and V. A. Zinoviev, *On the cosets of weight* 4 *of binary BCH codes with minimum distance* 8 *and exponential sums*, Probl. Inf. Transm., 41 (2005), pp. 331–348.
[5] P. Charpin, T. Helleseth, and V. A. Zinoviev, *The divisibility modulo* 24 *of Kloosterman sums on* $GF(2^m)$*,* m *odd*, J. Combin. Theory Ser. A, 114 (2007), pp. 322–338.
[6] P. Charpin, T. Helleseth, and V. A. Zinoviev, *The coset distribution of triple-error-correcting binary primitive BCH codes*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1727–1732.
[7] P. Charpin, T. Helleseth, and V. A. Zinoviev, *Propagation characteristics of* $x \mapsto x^{-1}$ *and Kloosterman sums*, Finite Fields Appl., 13 (2007), pp. 366–381.
[8] G. van der Geer and M. van der Vlugt, *The coset weight distributions of certain BCH codes and a family of curves*, Enseign. Math., 48 (2002), pp. 3–21.
[9] T. Helleseth, *All binary* 3*-error-correcting BCH codes of length* $2^m - 1$ *have covering radius* 5, IEEE Trans. Inform. Theory, 24 (1978), pp. 257–258.
[10] T. Helleseth and V. A. Zinoviev, *On* Z_4*-linear Goethals codes and Kloosterman sums*, Des. Codes Cryptogr., 17 (1999), pp. 269–288.
[11] T. Helleseth and V. A. Zinoviev, *On coset weight distributions of the* Z_4*-linear Goethals codes*, IEEE Trans. Inform. Theory, 47 (2001), pp. 1758–1772,
[12] T. Helleseth and V. A. Zinoviev, *On a new identity for Kloosterman sums and nonlinear system of equations over finite fields of characteristic* 2, Discrete Math., 274 (2004), pp. 109–124.
[13] P. V. Kumar, T. Helleseth, R. A. Calderbank, and R. A. Hammons, *Large families of quaternary sequences with low correlation*, IEEE Trans. Inform. Theory, 42 (1996), pp. 579–592.
[14] G. Lachaud and J. Wolfmann, *The weights of the orthogonals of the extended quadratic binary Goppa codes*, IEEE Trans. Inform. Theory, 36 (1990), pp. 686–692.
[15] R. Lidl and H. Niederreiter, *Finite Fields*, Encyclopedia Math. Appl. 20, Addison-Wesley, Reading, MA, 1983.
[16] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error Correcting Codes*, North-Holland, Amsterdam, 1986.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 79–95

ON THE DISTRIBUTION OF BOOLEAN FUNCTION NONLINEARITY*

SIMON LITSYN† AND ALEXANDER SHPUNT‡

Abstract. Nonlinearity is the number of bits which must change in the truth table of a Boolean function to reach the closest affine function. It may be expressed through the maximum of the absolute value of a component in the function's Walsh–Hadamard transform. Concentration of nonlinearity is proved. The derived bounds on the concentration point and tails of the distribution are tighter than the earlier known ones.

Key words. Boolean functions, concentration of nonlinearity, Walsh–Hadamard transform, binomial sums, tails of binomial distribution, second moment method

AMS subject classifications. Primary, 94C10; Secondary, 11T71, 94A60, 06E30

DOI. 10.1137/060665361

1. Introduction and definitions. Nonlinearity is the number of bits which must change in the truth table of a Boolean function to reach the closest affine function. The notion has multiple applications in coding theory and cryptography. For example, nonlinearity can be used as a measure of strength of cryptosystems. It is particularly useful to quantify the strength of invertible substitution tables when predefined tables are a part of a cipher definition; see, e.g., [3]. Nonlinearity plays an important role in relation to the covering radius of the first-order Reed–Muller codes; see [5] and the references therein. Namely, the most nonlinear (bent) functions correspond to the farthest-off, from the code, vectors in the ambient space. One of the widely addressed problems in this context is enumeration of Boolean functions according to their nonlinearity. This has been successfully accomplished by Berlekamp and Welch [2] in the case of up to five variables and by Maiorana [9] for six variables. However, exact enumeration for a greater number of variables seems to be intractable. Therefore, estimates on the distribution of nonlinearity become relevant. This was attempted by Carlet [3, 4], Olejár and Stanek [10], Rodier [11, 12], and Wu [16]. Especially interesting was a recent result of Rodier [13] where by using a method from harmonic analysis due to Halász [6] he proved a concentration of the nonlinearity. In this paper we further develop this theme by proving tighter results for the concentration point and the tails of the distribution. Moreover, though quite technical, the developed approach is basically the second moment method (see, e.g., [1]) and is conceptually much simpler than the Halász approach.

Let $f = f(x_1, x_2, \ldots, x_m)$ and $h = h(x_1, x_2, \ldots, x_m)$, $x_i \in \{0, 1\}$, $i = 1, 2, \ldots, m$, be Boolean functions taking on values from $\{0, 1\}$. The (Hamming) distance between two functions, $d(f, h)$, is the number of strings $x_1, \ldots, x_m$, for which $f \neq h$. Nonlinearity of f, $nl(f)$ is

$$nl(f) = \min_h d(f, h),$$

*Received by the editors July 18, 2006; accepted for publication (in revised form) June 17, 2008; published electronically October 24, 2008.
http://www.siam.org/journals/sidma/23-1/66536.html
†Department of Electrical Engineering-Systems, Tel Aviv University, Ramat Aviv, 69978, Israel (litsyn@eng.tau.ac.il). This author was supported in part by ISF grant 533-06.
‡Department of Physics, Massachusetts Institute of Technology, 77 Mass. Ave., Cambridge, MA 02139 (ashpunt@mit.edu).

where the minimum is taken over all affine functions h,

$$h = a_0 + a_1 x_1 + \cdots + a_m x_m, \quad a_i \in \{0,1\},\ i = 0,1,\ldots,m.$$

The spectral amplitude of f, $S(f)$ is

$$S(f) = \max_{v \in \{0,1\}^m} \left| \sum_{u \in \{0,1\}^m} (-1)^{f(u)+v\cdot u} \right|,$$

where $v \cdot u$ is the usual dot product. In other words, the spectral amplitude of f is just the maximum absolute value of a component in the Walsh–Hadamard transform of f. Indeed, let $n = 2^m$, $\mathcal{M}$ be the $n \times n$ Walsh–Hadamard matrix, its rows being $\mathcal{M}_i$, and entries $\mathcal{M}_{i,j}$, $i,j = 0,\ldots,n-1$. We have

$$\mathcal{M}_{i,j} = (-1)^{i_0 j_0 + i_1 j_1 + \cdots + i_{m-1} j_{m-1}},$$

where $(i_0, i_1, \ldots, i_{m-1})_2$ and $(j_0, j_1, \ldots, j_{m-1})_2$ are the binary expansions of i and j correspondingly (indeed, $M_{i,j}$ is the ith character evaluated at point j). Then, denoting

$$f_j^* = (-1)^{f(j_0,\ldots,j_{m-1})}$$

for the binary expansion $(j_0,\ldots,j_{m-1})_2$ of j, and

$$\mathcal{M}_i^*(f) = \sum_{j=1}^{n} f_j^* \cdot \mathcal{M}_{i,j},$$

we conclude that

$$S(f) = \max_{i=0,\ldots,n-1} |\mathcal{M}_i^*(f)|. \tag{1.1}$$

Notice that $\mathcal{M}_i$ is the evaluation of the linear function

$$L_i(x_1,\ldots,x_m) = i_0 x_1 + i_1 x_2 + \cdots + i_{m-1} x_{m-1},$$

namely,

$$\mathcal{M}_{i,j} = (-1)^{L_i(j_0,\ldots,j_{m-1})}.$$

Moreover, all possible linear functions are presented as rows of $\mathcal{M}$. Therefore, since

$$\mathcal{M}_i^*(f) = n - 2d(f, L_i),$$

there is a simple relation between the spectral amplitude and nonlinearity,

$$nl(f) = 2^{m-1} - \frac{1}{2} S(f).$$

It is a simple corollary of the Parseval identity,

$$\sum_{i=0}^{n-1} (\mathcal{M}_i^*(f))^2 = n^2,$$

that

$$\max_i |\mathcal{M}_i^*(f)| \geq \sqrt{n} = 2^{\frac{m}{2}}.$$

Therefore, for any Boolean function in m variables,

$$nl(f) \leq 2^{m-1} - 2^{\frac{m}{2}-1},$$

and this bound is achieved only when m is even.

The introduced notions can be straightforwardly extended to general Hadamard matrices; see, e.g., [7, 15]. Indeed, the Hadamard $n \times n$ matrices $\mathcal{H}$ consist of ± 1's and satisfy

$$\mathcal{H}\mathcal{H}^T = \mathcal{I},$$

and the Walsh–Hadamard matrices, described earlier, constitute a subclass of Hadamard matrices for $n = 2^m$. A well-known conjecture states that such matrices exist for $n = 2$ and all natural n divisible by 4. Similar to the Walsh–Hadamard case, for a ± 1-vector w of length n define its spectral amplitude and nonlinearity as

$$S(v) = \max_{i=0,1,\ldots,n-1} |(\mathcal{H}v^T)_i|,$$

and

$$nl(v) = \frac{n}{2} - \frac{1}{2}S(v),$$

correspondingly. To simplify, in what follows we restrict our claims to Boolean functions; however, we keep in mind that a more general situation is under consideration, and thus n can take on arbitrary positive integer values for which Hadamard matrices of size n exist.

Now we are in a position to restate the result of Rodier in a more rigorous form than it appears in [13]. We will use a probabilistic terminology. Namely, we will be considering probabilities of events in the ensemble of $2^n = 2^{2^m}$ equiprobable Boolean functions in m variables. The same can be undertaken in the general Hadamard case, if one deals with the ensemble of 2^n equiprobable vectors of length n.

THEOREM 1.1 (Rodier–Halász).

$$\Pr\left(|S(f)| > \sqrt{2n(\ln n + 5.4 \ln\ln n(1 + o(1)))}\ \right) = O\left(\frac{1}{\ln^4 n}\right),$$

$$\Pr\left(|S(f)| < \sqrt{2n(\ln n - 7 \ln\ln n(1 + o(1)))}\ \right) = O\left(\frac{1}{\ln^4 n}\right).$$

A comment is in order here. Indeed the theorem claims concentration of the spectral amplitude around $\sqrt{2n \ln n}$. Notice that the summand $\text{const} \cdot \ln\ln n$ is inevitable in the Halász approach and cannot be removed or decreased to some slower growing in n function.

Before we state the results of the present paper, let us briefly examine the problem from a geometrical point of view. Note that the functions f with $nl(f) \leq \rho$ are those for which there exists an i such that either $d(f, L_i) \leq \rho$ or $d(f, \overline{L_i}) \leq \rho$, where $\overline{L_i}$ is the 1's complement of L_i. Therefore, the problem of computing the number of

functions with nonlinearity $\le \rho$ reduces to computing the volume $\mathcal{V}(\rho)$ of the union of $2n$ Hamming spheres, each of radius ρ, centered at $\pm\mathcal{M}_i$, $i = 1, 2, \ldots, n$. The center of each sphere is at the Hamming distance n from one other sphere and at distance $n/2$ from the remaining $2n-2$ spheres. The volume of intersection of any two of the above nonantipodal spheres can easily be shown to be (see, e.g., [5, section 2.4, "Hamming spheres"])

$$\sum_{i,j=0}^{\rho} \binom{\frac{n}{2}}{\frac{i+j}{2}-\frac{n}{4}} \binom{\frac{n}{2}}{\frac{i-j}{2}+\frac{n}{4}} \le n^2 \cdot \binom{\frac{n}{2}}{\frac{n}{4}} \binom{\frac{n}{2}}{\rho-\frac{n}{4}}$$

for $\rho \ge n/4$ and zero for $\rho < n/4$. The number of intersecting sphere pairs is $< 2n^2$, and consequently, the total volume of intersections is

$$V_\cap < 2n^4 \cdot \binom{\frac{n}{2}}{\frac{n}{4}} \binom{\frac{n}{2}}{\rho-\frac{n}{4}}.$$

The volume of one such sphere is of course $> \binom{n}{\rho}$, and consequently, the volume of $2n$ spheres is

$$V_\Sigma > 2n \binom{n}{\rho}.$$

By the inclusion-exclusion principle,

$$V_\Sigma - V_\cap \le \mathcal{V}(\rho) \le V_\Sigma.$$

For all n, such that $\rho_0 = n/2 - \lceil\sqrt{2n \ln n}\rceil > n/4$, it can be seen that $2n \cdot \binom{n}{\rho_0}$ is much greater than $2n^4 \cdot \binom{n/2}{n/4}\binom{n/2}{\rho_0 - n/4}$, in fact, asymptotically so (for growing n). It should be noted here that for $n < 164$, we have $\rho_0 < n/4$, and the spheres with $0 \le \rho \le \rho_0$ have empty intersection. For the sake of simplicity, we henceforth restrict our treatment to $n \ge 164$.

Observe that $\binom{n}{\rho}/\binom{n/2}{\rho-n/4}$ decreases with growing ρ for $n/4 \le \rho < n/2 - 1/2$; therefore, we conclude that for $0 \le \rho \le \rho_0$, $\mathcal{V}(\rho)$ is asymptotically given by the *union bound* V_Σ (volume of one Hamming sphere times the number of spheres, n). However, one of the conclusions of this paper is that the union bound, in fact, is asymptotically tight for

$$\frac{n}{2} - \frac{n}{4} \le \rho < \frac{n}{2} - \frac{\sqrt{2n(\ln n - 0.5 \ln\ln n)}}{2},$$

which together with the previous statement provides the asymptotical exactness of the union bound on the whole upper tail of the Boolean spectral amplitude distribution, i.e.,

$$0 \le \rho < \frac{n}{2} - \frac{\sqrt{2n(\ln n - 0.5 \ln\ln n)}}{2}.$$

To achieve this we use the second moment method. Namely, we estimate the probability of each $|\mathcal{M}_i^*|$ in (1.1) to exceed some threshold. Next, for arbitrary i and j, we bound the probability that $|\mathcal{M}_i^*|$ and $|\mathcal{M}_j^*|$ are simultaneously above the threshold. This is followed by application of the Chebyshev inequality. Surprisingly

enough this allows one to tighten the bounds on concentration of the nonlinearity. We prove that the spectral amplitude is concentrated around $\sqrt{2n\left(\ln n - \frac{1}{2}\ln\ln n\right)}$. Moreover, we are able to derive explicit bounds on the tails of the distribution. A particular case of our main result is the following theorem.

THEOREM 1.2. *Let $n \geq 164$. Then the following holds true. For $-0.5\ln\ln n + 0.125 < \delta(n) < -\ln n + n/8$,*

$$\Pr\left(|S(f)| \geq \sqrt{2n(\ln n + \delta(n))}\right) \leq \left(\pi(\ln n + \delta(n))\right)^{-\frac{1}{2}} \cdot e^{-\delta(n)} \cdot (1 + o(1)).$$

Moreover, for $-0.5\ln\ln n + 0.125 < \delta(n) < o(\sqrt{n})$,

$$\Pr\left(|S(f)| \geq \sqrt{2n(\ln n + \delta(n))}\right) = \left(\pi(\ln n + \delta(n))\right)^{-\frac{1}{2}} \cdot e^{-\delta(n)} \cdot (1 + o(1)).$$

Finally, for the lower tail, $-\ln n + 0.5\ln\ln n < \delta(n) < -0.5\ln\ln n$,

$$\Pr\left(|S(f)| \geq \sqrt{2n(\ln n + \delta(n))}\right) \leq \left(\pi(\ln n + \delta(n))\right)^{\frac{1}{2}} \cdot e^{-\delta(n)} \cdot (1 + o(1)).$$

Comparing this bound with the one of Rodier, we notice that even in the case of $\delta(n)$ being of order $\ln\ln n$ our bounds are tighter.

This paper is organized as follows. Section 2 gives tight bounds for sums of binomial coefficients, appearing in all estimates of the binomial distribution tails. To the best of our knowledge, these are tighter than has been known before and can be useful in other research. In section 3, the tails of joint probability of two (dependent) events are estimated. In section 4 the exact asymptotics for the upper tail of the distribution of Boolean function nonlinearity is given, whereas section 5 provides a tight upper bound on the lower tail of this distribution. We also elaborate on the fact that the concentration point of the above distribution is localized more exactly than known before.

The following notations are used throughout. The Gaussian (normal) distribution with mean μ and standard deviation σ is denoted by $N(\mu, \sigma)$. The standard normal cumulative distribution function (CDF) is denoted by $F_G(x)$, and its complementary (tails) function by $P_G(x)$. Explicitly,

$$P_G(x) \equiv 1 - F_G(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt = \frac{1}{\sqrt{\pi}}\int_{x/\sqrt{2}}^\infty e^{-t^2}dt.$$

2. Bounds on binomial distribution tails. We start with auxiliary results concerning binomial coefficients. The numerical factors in front of the subleading terms are not optimal but were rather chosen to make simple expressions.

LEMMA 2.1. *For $0 < \epsilon_1 < \sqrt{3/32}$ and all n, such that $n \cdot \left(\frac{1}{2} - \epsilon_1\right)$ is an integer,*

$$2^{-n} \cdot \binom{n}{n \cdot (\frac{1}{2} - \epsilon_1)} \leq (1 + \varsigma_1) \cdot \sqrt{\frac{2}{\pi n}} \cdot e^{-2n\epsilon_1^2}, \quad \varsigma_1 < 3\epsilon_1^2. \tag{2.1}$$

Moreover, for $0 < \epsilon_1 < (2n)^{-1/4}$ and $n \geq 164$, such that $n \cdot \left(\frac{1}{2} - \epsilon_1\right)$ is an integer,

$$2^{-n} \cdot \binom{n}{n \cdot (\frac{1}{2} - \epsilon_1)} \geq (1 - \varsigma_2) \cdot \sqrt{\frac{2}{\pi n}} \cdot e^{-2n\epsilon_1^2}, \quad \varsigma_2 < \frac{3}{2}n\epsilon_1^4 + \frac{1}{2n}. \tag{2.2}$$

Proof. See section A.1 of the appendix. ∎

Let us derive explicit error bounds for the approximation of sums of binomial coefficients by a Gaussian CDF.

LEMMA 2.2. *Let*

$$S(n,d)=\sum_{k=\frac{d}{2}}^{\frac{n}{2}}\binom{n}{n\cdot(\frac{1}{2}-\frac{k}{n})}.$$

Then, for $n\geq 164$ *and* $\sqrt{n\ln\ln n}<d<n/2$, *the following inequalities hold:*

$$2^{-n}\cdot S(n,d)\leq(1+\varsigma_3)\cdot P_G\left(\frac{d}{\sqrt{n}}\right),\quad \varsigma_3<\frac{7d}{n}, \tag{2.3}$$

and for $\sqrt{n\ln\ln n}<d<(2n)^{3/4}$,

$$2^{-n}\cdot S(n,d)\geq(1-\varsigma_4)\cdot P_G\left(\frac{d}{\sqrt{n}}\right)-e^{-\sqrt{n/32}},\quad \varsigma_4<\frac{1}{2n}+\frac{5d^4}{n^3}. \tag{2.4}$$

Proof. See section A.2 of the appendix. ∎

COROLLARY 2.3. *Under the appropriate conditions of Lemma* 2.2, *for*

$$d=\sqrt{2n(\ln n+\delta(n))},$$

$$2^{-n}\cdot S(n,d)\leq\frac{1}{n}\cdot\frac{e^{-\delta(n)}}{\sqrt{2\pi d^2/n}}\cdot(1+\varsigma_3). \tag{2.5}$$

Moreover,

$$2^{-n}\cdot S(n,d)\geq\frac{1}{n}\cdot\frac{e^{-\delta(n)}}{\sqrt{2\pi d^2/n}}\cdot(1-\varsigma_4)\cdot(1-\varsigma_5)-e^{-\sqrt{n/32}},\quad \varsigma_5\leq\frac{n}{d^2}. \tag{2.6}$$

Proof. The proof follows trivially from Lemma 2.2. ∎

3. The probability of intersection. Let us define the events

$$A_i(d)=\{|\mathcal{M}_i^*|>d\},\ i=1,\ldots,n,$$

and estimate the probability of $A_{i_1}(d)\wedge A_{i_2}(d)$, $i_1\neq i_2$. Note that by construction,

$$P\left(M_i^*=a\right)=2^{-n}\cdot\binom{n}{n\cdot(\frac{1}{2}-\frac{a}{2n})}.$$

Likewise,

$$P\left(M_i^*\geq a\right)=2^{-n}\cdot S(n,a),$$

and we can use the results of section 2 to bound the probabilities of events $A_i(d)$ and their intersections.

LEMMA 3.1. *For* $i_1\neq i_2$, *under the conditions of Lemma* 2.2,

$$P\left(A_{i_1}(d)\wedge A_{i_2}(d)\right)\leq 4\cdot(1+\varsigma_6)\cdot\left(P_G\left(\frac{d}{\sqrt{n}}\right)\right)^2, \tag{3.1}$$

where $\varsigma_6=o(n^{-1/4})$. *(We omit the cumbersome explicit expression which can be easily developed.)*

Proof. Recall that for some Boolean function f,

$$\mathcal{M}^*_{i_1}(f) = \sum_{j=1}^{n} f^*_j \cdot \mathcal{M}_{i_1,j}, \quad \mathcal{M}^*_{i_2}(f) = \sum_{j=1}^{n} f^*_j \cdot \mathcal{M}_{i_2,j}.$$

Clearly, for $i_1 \neq i_2$,

$$P(A_{i_1}(d) \wedge A_{i_2}(d)) = 4P\left(\mathcal{M}^*_{i_1} > d \wedge \mathcal{M}^*_{i_2} > d\right). \tag{3.2}$$

Note that $\mathcal{M}_{i_1}, \mathcal{M}_{i_2}$ are two rows of some Walsh–Hadamard matrix; consequently, they are orthogonal $\{-1,+1\}^n$ vectors, and

$$E\{\mathcal{M}^*_{i_1} \cdot \mathcal{M}^*_{i_2}\} = \langle \mathcal{M}_{i_1}, \mathcal{M}_{i_2} \rangle = 0. \tag{3.3}$$

Hence, their vector half-sum, $\Sigma_{i_1 i_2} \equiv \frac{\mathcal{M}_{i_1} + \mathcal{M}_{i_2}}{2}$ has $\frac{n}{2}$ zeros and $\frac{n}{2}$ nonzero positions, where $\mathcal{M}_{i_1}$ and $\mathcal{M}_{i_2}$ are different or equal accordingly. Similarly, their vector half-difference $\Delta_{i_1 i_2} \equiv \frac{\mathcal{M}_{i_1} - \mathcal{M}_{i_2}}{2}$ has complementary $\frac{n}{2}$ zeros and $\frac{n}{2}$ nonzero positions, where $\mathcal{M}_{i_1}$ and $\mathcal{M}_{i_2}$ are equal or different accordingly. Consequently, events $\{\Sigma_{i_1 i_2} = a\}$ and $\{\Delta_{i_1 i_2} = b\}$ are independent for any a, b since they depend on mutually disjoint sets of independent random variables.

In order to establish the relation (3.5), we use the following trick. Assume for the sake of discussion that f^*_j are independently and identically distributed (i.i.d.) Gaussian random variables (rvs) and $f^*_j \sim N(0,1)$, $j = 1, 2, \ldots, n$. It follows from (3.3) that $\mathcal{M}^*_{i_1}(f)$ and $\mathcal{M}^*_{i_2}(f)$ are then also i.i.d. Gaussian rvs (for Gaussian rvs, zero cross-correlation implies independence), with $\mathcal{M}^*_{i_1}(f), \mathcal{M}^*_{i_2}(f) \sim N(0, \sqrt{n})$.

Consequently,

$$P\left(\mathcal{M}^*_{i_1} > d \wedge \mathcal{M}^*_{i_2} > d\right) = P\left(\mathcal{M}^*_{i_1} > d\right) \cdot P\left(\mathcal{M}^*_{i_2} > d\right) = P^2_G(d/\sqrt{n}). \tag{3.4}$$

On the other hand, since $\Sigma_{i_1 i_2}, \Delta_{i_1 i_2}$ are i.i.d. $\sim N(0, \sqrt{\frac{n}{2}})$,

$$P\left(\mathcal{M}^*_{i_1} > d \wedge \mathcal{M}^*_{i_2} > d\right) = \int_d^\infty P\left(\Sigma_{i_1 i_2} = x\right) dx \int_{-(x-d)}^{x-d} P\left(\Delta_{i_1 i_2} = t\right) dt$$

$$= \frac{1}{\pi} \int_{\frac{d}{\sqrt{n}}}^\infty e^{-x^2} dx \int_{-\frac{x-d}{\sqrt{n}}}^{\frac{x-d}{\sqrt{n}}} e^{-t^2} dt.$$

Consequently, we have established that

$$\frac{1}{\pi} \int_{\frac{d}{\sqrt{n}}}^\infty e^{-x^2} dx \int_{-\frac{x-d}{\sqrt{n}}}^{\frac{x-d}{\sqrt{n}}} e^{-t^2} dt = P^2_G(d/\sqrt{n}). \tag{3.5}$$

Armed with (3.5), let us now consider our case where f^*_j are binary random variables. Since $\Sigma_{i_1 i_2}$ and $\Delta_{i_1 i_2}$ are i.i.d. (but not Gaussian in this case), we can write

$$P(d) \equiv P\left(\mathcal{M}^*_{i_1} > d \wedge \mathcal{M}^*_{i_2} > d\right) = \sum_{a=\frac{d}{2}}^{\frac{n}{4}} P\left(\Sigma_{i_1 i_2} = 2a\right) \cdot \sum_{b=-\left(a-\frac{d}{2}\right)}^{a-\frac{d}{2}} P\left(\Delta_{i_1 i_2} = 2b\right).$$

Since $\Sigma_{i_1 i_2}$ and $\Delta_{i_1 i_2}$ are disjoint (independent) sums of $\frac{n}{2}$ (binary, $\{+1,-1\}$) independent random variables, $P(\Sigma_{i_1 i_2} = 2a)$ is the probability that the excess of $+1$'s in $\Sigma_{i_1 i_2}$ will be a; same for $\Delta_{i_1 i_2}$ and b. Hence

$$P(\Sigma_{i_1 i_2} = 2a) = 2^{-n/2} \cdot \binom{\frac{n}{2}}{\frac{n}{2}\left(\frac{1}{2} - \frac{2a}{n}\right)} \equiv P_1(a),$$

$$P(\Delta_{i_1 i_2} = 2b) = 2^{-n/2} \cdot \binom{\frac{n}{2}}{\frac{n}{2}\left(\frac{1}{2} - \frac{2b}{n}\right)} \equiv P_2(b).$$

For $a \le a_0 = \frac{n^{3/4}}{4\sqrt{\ln \ln n}}$, by (A.4), and using $\sqrt{1+x} < 1 + \sqrt{x}$ for $x > 0$,

$$P_1(a) \le \frac{2}{\sqrt{\pi n}} \cdot \left(1 + \frac{1}{(n^{1/2} \ln \ln n - 1)^{1/2}}\right) \cdot e^{-4a^2/n}. \tag{3.6}$$

For $|a| > a_0$,

(3.7)

$$P_1(a) \le P_1(a_0) \le \frac{2}{\sqrt{\pi n}} \cdot \left(1 + \frac{1}{(n^{1/2} \ln \ln n - 1)^{1/2}}\right) \cdot e^{-\frac{\sqrt{n}}{4 \ln \ln n}} < \sqrt{\frac{2}{n}} \cdot e^{-\frac{\sqrt{n}}{4 \ln \ln n}}.$$

Consequently,

$$P(d) = \sum_{a=\frac{d}{2}}^{\frac{n}{4}} P_1(a) \cdot \sum_{b=-\left(a-\frac{d}{2}\right)}^{a-\frac{d}{2}} P_2(b) \le \sum_{a=\frac{d}{2}}^{a_0} P_1(a) \cdot \sum_{b=-\left(a-\frac{d}{2}\right)}^{a-\frac{d}{2}} P_2(b) + e^{-\frac{\sqrt{n}}{20 \ln \ln n}}$$

$$\le \left(1 + o(n^{-1/4})\right) \cdot \frac{4}{\pi n} \cdot \sum_{a=\frac{d}{2}}^{\frac{n}{4}} e^{-4a^2/n} \cdot \sum_{b=-(a-\frac{d}{2})}^{a-\frac{d}{2}} e^{-4b^2/n} + e^{-\frac{\sqrt{n}}{20 \ln \ln n}}.$$

From here (shown in detail in subsection A.3), we see that

$$P(d) \le \left(1 + o(n^{-1/4})\right) \cdot \frac{1}{\pi} \int_{\frac{d}{\sqrt{n}}}^{\infty} e^{-x^2} dx \int_{-\frac{2x-d}{\sqrt{n}}}^{\frac{2x-d}{\sqrt{n}}} e^{-z^2} dz. \tag{3.8}$$

Finally, by using (3.5) we have obtained

$$P(d) \le \left(1 + o(n^{-1/4})\right) \cdot \left(P_G\left(\frac{d}{\sqrt{n}}\right)\right)^2. \tag{3.9}$$ ∎

4. The upper tail. The results of the previous sections enable us to derive an asymptotically exact result for the upper tail of the distribution of the spectral amplitude $S(f)$ of Boolean functions, as defined by (1.1).

THEOREM 4.1. *For $n \ge 164$, let*

$$d = \sqrt{2n(\ln n + \delta(n))}.$$

Then, for $\sqrt{2n(\ln n - 0.5 \ln \ln n + 0.125)} < d < n/2$,

$$\Pr\left(\max_{i=1,\dots,n} |\mathcal{M}_i^*| > d\right) \le \frac{e^{-\delta(n)}}{\sqrt{\pi d^2/(2n)}} \cdot \left(1 + \frac{7d}{n}\right). \tag{4.1}$$

Moreover, for $\sqrt{2n(\ln n - 0.5\ln\ln n + 0.125)} < d < o\left(n^{3/4}\right)$,

$$\Pr\left(\max_{i=1,\ldots,n} |\mathcal{M}_i^*| > d\right) \geq \frac{e^{-\delta(n)}}{\sqrt{\pi d^2/(2n)}} \cdot (1 - o(1)). \tag{4.2}$$

Proof. Using the definition of $A_i(d)$ from section 3,

$$\Pr\left(\max_{i=1,\ldots,n} |\mathcal{M}_i^*| > d\right) = 2 \cdot \Pr\left(\max_{i=1,\ldots,n} \mathcal{M}_i^* > d\right) = \Pr\left(\bigcup_{i=1}^{n} A_i(d)\right).$$

To prove (4.1), use

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_i \Pr(A_i)$$

and Corollary 2.3.
For (4.2) use the inclusion-exclusion principle,

$$\Pr\left(\bigcup_{i=1}^{n} A_i(d)\right) \geq \sum_i \Pr(A_i(d)) - \sum_{\substack{i_1,\, i_2 \\ i_1 > i_2}} \Pr(A_{i_1}(d) \wedge A_{i_2}(d)).$$

From Corollary 2.3, for $d = o(n^{3/4})$ and $i = 1, \ldots, n$,

$$\Pr(A_i(d)) = 2\Pr(\mathcal{M}_i^* > d) \geq \frac{1}{n} \cdot \frac{e^{-\delta(n)}}{\sqrt{\pi(\ln n + \delta(n))}} (1 - o(1)). \tag{4.3}$$

From Lemma 3.1,

$$\begin{aligned}\Pr(A_{i_1}(d) \wedge A_{i_2}(d)) = 4P\left(\mathcal{M}_i^* > d \wedge \mathcal{M}_j^* > d\right) &\leq 4 \cdot (1 + o(1)) \cdot \left(P_G\left(\frac{d}{\sqrt{n}}\right)\right)^2 \\ &\leq \frac{2}{n^2} \cdot \frac{e^{-2\delta(n)}}{\pi d^2/(2n)} \cdot (1 + o(1)),\end{aligned}$$

where we once again used $P_G(x) \leq e^{-x^2/2}/(\sqrt{2\pi}x)$. From here, it follows immediately that

$$\Pr\left(\bigcup_{i=1}^{n} A_i(d)\right) \geq \frac{e^{-\delta(n)}}{\sqrt{\pi d^2/(2n)}} \cdot \left(1 - \frac{e^{-\delta(n)}}{\sqrt{\pi d^2/(2n)}}\right) \cdot (1 - o(1)).$$

Note that the condition

$$\frac{e^{-\delta(n)}}{\sqrt{\pi d^2/(2n)}} < \frac{1}{2}$$

gives us the lower bound for the value of $d > \sqrt{2n(\ln n - 0.5(\ln\ln n - \ln(4/\pi)))}$, for which the theorem holds. ∎

Remark 4.2. From Theorem 4.1 we see that for $\delta(n)$ growing with n, the sum of probabilities of all pairwise intersections

$$A_{i_1}\left(\sqrt{2n(\ln n + \delta(n))}\right) \wedge A_{i_2}\left(\sqrt{2n(\ln n + \delta(n))}\right)$$

is asymptotically smaller than the union bound

$$\frac{e^{-\delta(n)}}{\sqrt{\pi(\ln n+\delta(n))}}.$$

From Theorem 4.1 we see that the distribution is indeed concentrated around a point to the left of $\sqrt{2n(\ln n-0.5\ln\ln n)}$. Let us state this in the following corollary.

COROLLARY 4.3. *For* $0<\beta<\frac{(\sqrt{n}/\ln\ln n)-\ln n}{\ln\ln n}+0.5$,

$$\Pr\left(\max_{i=1,\dots,n}|\mathcal{M}_i^*|>\sqrt{2n(\ln n-0.5\ln\ln n+\beta\ln\ln n+0.125)}\right)=O\left(\frac{1}{\ln^\beta n}\right). \tag{4.4}$$

Proof. Take

$$\delta(n)=(\beta-0.5)\ln\ln n. \tag{4.5}$$

Substitute (4.5) into (4.1) and (4.2) to get

$$\Pr\left(\max_{i=1,\dots,n}|\mathcal{M}_i^*|>\sqrt{2n(\ln n-0.5\ln\ln n+\beta\ln\ln n+0.125)}\right)$$
$$=\frac{(\ln n)^{-\beta}}{\sqrt{\pi}}(1+o(1)).\qquad\square$$

From Corollary 4.3, we have, for example,

$$\Pr\left(\max_{i=1,\dots,n}|\mathcal{M}_i^*|>\sqrt{2n(\ln n+3.5\ln\ln n+0.125)}\right)=O\left(\frac{1}{\ln^4 n}\right).$$

5. The lower tail. Results obtained in the previous sections enable us to tightly bound the lower tail of $S(f)$, using a variation of the second moment method.

THEOREM 5.1. *For* $n\ge 164$ *and*

$$d=\sqrt{2n(\ln n-\delta(n))},\quad \sqrt{n\ln\ln n}<d<\sqrt{2n(\ln n-0.5\ln\ln n)},$$
$$\Pr\left(\max_{i=1,\dots,n}|\mathcal{M}_i^*|\le d\right)\le\sqrt{\pi d^2/(2n)}\cdot e^{-\delta(n)}\cdot(1+o(1)). \tag{5.1}$$

Proof. For the events $A_i(d)\equiv\{|\mathcal{M}_i^*|>d\}$, and their indicators

$$I_{A_i(d)}(\omega)=[\omega\in A_i(d)],\ i=1,2,\dots,n,$$

let

$$A=\sum_{i=1}^{n}I_{A_i(d)}.$$

Then,

$$\Pr\left(\max_{i=1,\dots,n}|\mathcal{M}_i^*|\le d\right)=\Pr(A=0).$$

By linearity of the expectation,

$$E\{A\}=2\cdot n\cdot\Pr\left(\mathcal{M}_{\mathrm{i}}^*>\mathrm{d}\right).$$

By Lemma 2.2,

$$\Pr\left(\mathcal{M}_{\mathrm{i}}^{*} > \mathrm{d}\right) \geq \mathrm{P_G}\left(\frac{\mathrm{d}}{\sqrt{\mathrm{n}}}\right)(1-\varsigma_4) - \mathrm{e}^{-\sqrt{\mathrm{n}/32}}. \tag{5.2}$$

Furthermore,

$$E\{A^2\} = \sum_{i_1=1,i_2=1}^{n} E\{I_{A_{i_1}(d)} \cdot I_{A_{i_2}(d)}\} = \sum_{i_1=1,i_2=1}^{n} \Pr\left(A_{i_1}(d) \wedge A_{i_2}(d)\right).$$

By Lemma 3.1,

$$E\{A^2\} \leq n(n-1)P\left(A_{i_1}(d) \wedge A_{i_2}(d)\right) + E\{A\} \leq n^2 P\left(A_{i_1}(d) \wedge A_{i_2}(d)\right) + E\{A\}$$

$$\leq 4n^2 \cdot \left(1 + o(n^{-1/4})\right) \cdot \left(P_G\left(\frac{d}{\sqrt{n}}\right)\right)^2 + E\{A\}.$$

Therefore

$$\mathrm{Var}\left(\mathrm{A}\right) = \mathrm{E}\{\mathrm{A}^2\} - (\mathrm{E}\{\mathrm{A}\})^2 \leq 4 \cdot n^2 \cdot o(n^{-1/4}) \left(P_G\left(\frac{d}{\sqrt{n}}\right)\right)^2 + E\{A\}.$$

Using the Chebyshev inequality,

$$P(A=0) \leq P\left(|A - E\{A\}| \geq E\{A\}\right) \leq \frac{\mathrm{Var}\{\mathrm{A}\}}{E^2\{A\}} \leq \frac{1}{E\{A\}}\left(1 + o(n^{-1/4}) \cdot E\{A\}\right).$$

Note, finally, that applying Corollary 2.3 to the expression for $E\{A\}$ obtained above, we have

$$E\{A\} \geq \frac{e^{\delta(n)}}{\sqrt{\pi(\ln n - \delta(n))}}\left(1 - o(1)\right). \qquad \square$$

From Theorems 4.1 and 5.1 we see that the distribution is indeed concentrated around $\sqrt{2n(\ln n - 0.5 \ln\ln n)}$. Also, we see that away from the concentration point, the distribution decays faster than established previously.

Corollary 5.2.

$$\Pr\left(\max_{i=1,\ldots,n} |\mathcal{M}_i^*| > \sqrt{2n(\ln n - 0.5\ln\ln n - \alpha \ln\ln n)}\right) = O\left(\frac{1}{\ln^{\alpha} n}\right). \tag{5.3}$$

From Corollary 5.2, we have, for example,

$$\Pr\left(\max_{i=1,\ldots,n} |\mathcal{M}_i^*| > \sqrt{2n(\ln n - 4.5\ln\ln n)}\right) = O\left(\frac{1}{\ln^4 n}\right).$$

6. Conclusions. Using only basic combinatorics, we provided the asymptotically exact upper tail and an upper bound on the lower tail for the distribution of nonlinearity of Boolean functions. These bounds yield a concentration of nonlinearity and are tighter than the earlier known ones. An open problem is estimating how tight our bound is on the lower tail. Notice that Spencer [14] provides a nonconstructive method guaranteeing an exponential number of functions with nonlinearity in the range of the lower tail. However, the bounds we were able to derive using this approach are very weak.

Appendix.

A.1. Proof of Lemma 2.1. Let us first show the upper bound. For $0 < \epsilon_1 < 1/2$ and any $n > 0$, we have

$$\frac{n!}{\left(n\cdot\left(\frac{1}{2}-\epsilon_1\right)\right)!\left(n\cdot\left(\frac{1}{2}+\epsilon_1\right)\right)!} \le \frac{1}{\sqrt{2\pi n\left(\frac{1}{4}-\epsilon_1^2\right)}}\cdot\frac{1}{\left(\frac{1}{2}+\epsilon_1\right)^{\left(\frac{1}{2}+\epsilon_1\right)n}\left(\frac{1}{2}-\epsilon_1\right)^{\left(\frac{1}{2}-\epsilon_1\right)n}},$$

where we have used (see, e.g., [8])

$$\sqrt{2\pi}\cdot n^{n+1/2}\cdot e^{-n+\frac{1}{12n}-\frac{1}{360n^3}} < n! < \sqrt{2\pi}\cdot n^{n+1/2}\cdot e^{-n+\frac{1}{12n}}. \tag{A.1}$$

Therefore, for $0 < \epsilon_1 < \frac{1}{2}$ and any $n > 0$,

$$\binom{n}{n\cdot\left(\frac{1}{2}-\epsilon_1\right)} \le \frac{1}{\sqrt{2\pi n\left(\frac{1}{4}-\epsilon_1^2\right)}}\cdot e^{nH_e\left(\frac{1}{2}-\epsilon_1\right)}, \tag{A.2}$$

where $H_e(x) \equiv -x\ln x-(1-x)\ln(1-x)$ stands for the natural entropy function.

Using also

$$H_e\left(\frac{1}{2}-\epsilon_1\right) \le \ln 2 - 2\epsilon_1^2 \quad \text{for } 0<\epsilon_1<\frac{1}{2}, \tag{A.3}$$

and

$$\frac{1}{\sqrt{1-x}} \le 1+\frac{3}{4}x \quad \text{for } 0\le x\le\frac{3}{8},$$

we have ($\epsilon_1^2 \le 3/32$)

$$\binom{n}{n\cdot\left(\frac{1}{2}-\epsilon_1\right)} \le \frac{1}{\sqrt{2\pi n\left(\frac{1}{4}-\epsilon_1^2\right)}}\cdot e^{n\ln 2-2n\epsilon_1^2} \le 2^n\cdot\left(1+3\epsilon_1^2\right)\cdot\sqrt{\frac{2}{\pi n}}\cdot e^{-2n\epsilon_1^2}. \tag{A.4}$$

Now to the lower bound. Using (A.1), we have for $0 < \epsilon_1 < \frac{1}{2}$ and any $n > 0$,

$$\frac{n!}{\left(n\cdot\left(\frac{1}{2}-\epsilon_1\right)\right)!\left(n\cdot\left(\frac{1}{2}+\epsilon_1\right)\right)!} \ge \frac{1}{\sqrt{2\pi n\left(\frac{1}{4}-\epsilon_1^2\right)}}\cdot\frac{e^{-\frac{1}{12\left(\frac{1}{2}+\epsilon_1\right)n}-\frac{1}{12\left(\frac{1}{2}-\epsilon_1\right)n}}}{\left(\frac{1}{2}+\epsilon_1\right)^{\left(\frac{1}{2}+\epsilon_1\right)n}\left(\frac{1}{2}-\epsilon_1\right)^{\left(\frac{1}{2}-\epsilon_1\right)n}}.$$

Therefore, for $0 < \epsilon_1 < \frac{1}{2}$ and any $n > 0$,

$$\binom{n}{n\cdot\left(\frac{1}{2}-\epsilon_1\right)} \ge \frac{1}{\sqrt{2\pi n\left(\frac{1}{4}-\epsilon_1^2\right)}}\cdot e^{n\cdot H_e\left(\frac{1}{2}-\epsilon_1\right)-1/\left(n\left(3-12\epsilon_1^2\right)\right)}.$$

For $0 \le \epsilon_1 \le (2n)^{-1/4}$ and $n \ge 164$,

$$H_e\left(\frac{1}{2}-\epsilon_1\right) \ge \ln 2 - 2\epsilon_1^2 - \frac{3}{2}\epsilon_1^4, \quad \text{and} \quad \frac{1}{n\left(3-12\epsilon_1^2\right)} \le \frac{1}{2n}.$$

Since $e^{-x} \ge 1-x$ for $x > 0$, we have

$$\binom{n}{n\cdot\left(\frac{1}{2}-\epsilon_1\right)} \ge 2^n\cdot\sqrt{\frac{2}{\pi n}}\cdot e^{-2n\epsilon_1^2}\cdot\left(1-\frac{3}{2}n\epsilon_1^4-\frac{1}{2n}\right). \tag{A.5}$$

□

A.2. Proof of Lemma 2.2. Let us first prove (2.3). Let

$$k_0 = \left\lfloor \sqrt{\frac{3}{32}} n \right\rfloor.$$

Then,

$$2^{-n} \cdot S(n,d) = \sum_{k=\frac{d}{2}}^{k_0} \binom{n}{n \cdot (\frac{1}{2} - \frac{k}{n})} + \sum_{k=k_0+1}^{\frac{n}{2}} \binom{n}{n \cdot (\frac{1}{2} - \frac{k}{n})} = S_1(n,d) + S_2(n,d).$$

Taking into account that the terms of $S_2(n,d)$ are monotonously decreasing, let us bound $S_2(n,d)$ from above by the product of the first (biggest) term and the number of terms in the sum, using Lemma 2.1,

$$S_2(n,d) < \frac{41}{64}\sqrt{\frac{2}{\pi}} \left(1 - \sqrt{\frac{3}{8}}\right) \cdot \sqrt{n} \cdot e^{-\frac{3n}{16}} < \sqrt{\frac{n}{4\pi}} \cdot e^{-3n/16}. \tag{A.6}$$

As for $S_1(n,d)$, we apply the upper bound of Lemma 2.1 to get

$$S_1(n,d) \le \sqrt{\frac{2}{\pi n}} \cdot \sum_{k=\frac{d}{2}}^{\infty} \left(1 + \frac{3k^2}{n^2}\right) \cdot e^{-2k^2/n}.$$

Bounding the sum with an integral, noting that for $d > \sqrt{n \ln \ln n}$ the integrands are monotonously decreasing functions of k, and recalling

$$P_G\left(\frac{d}{\sqrt{n}}\right) = \frac{1}{\sqrt{\pi}} \int_{\frac{d}{\sqrt{2n}}}^{\infty} e^{-z^2} dz,$$

we have

$$\sqrt{\frac{2}{\pi n}} \cdot \sum_{k=\frac{d}{2}}^{\infty} e^{-2k^2/n} < \sqrt{\frac{2}{\pi n}} \cdot e^{-\frac{d^2}{2n}} + P_G\left(\frac{d}{\sqrt{n}}\right),$$

$$\sqrt{\frac{2}{\pi n}} \cdot \frac{3}{2n} \cdot \sum_{k=\frac{d}{2}}^{\infty} \frac{2k^2}{n} e^{-2k^2/n} < \frac{3d(2d+1)}{\sqrt{32\pi n^5}} \cdot e^{-\frac{d^2}{2n}} + \frac{3}{4n} \cdot P_G\left(\frac{d}{\sqrt{n}}\right).$$

By assumption, we have

$$3d(2d+1)/\sqrt{32} < \sqrt{2}d^2, \quad (d/n)^2 + 1 < \sqrt{\pi/2},$$

and

$$\sqrt{\frac{2}{\pi n}} + \frac{3d(2d+1)}{\sqrt{32\pi n^5}} < \frac{1}{\sqrt{n}}.$$

Summing up and using (A.6),

$$2^{-n} \cdot S(n,d) < P_G\left(\frac{d}{\sqrt{n}}\right) \cdot \left(1 + \frac{3}{4n}\right) + \frac{e^{-\frac{d^2}{2n}}}{\sqrt{n}} + \sqrt{\frac{n}{4\pi}} \cdot e^{-3n/16}. \tag{A.7}$$

Noting that

$$\sqrt{\frac{n}{2\pi d^2}}\cdot e^{-\frac{d^2}{2n}}\cdot\left(1-\frac{n}{d^2}\right)\le P_G\left(\frac{d}{\sqrt{n}}\right)\le\sqrt{\frac{n}{2\pi d^2}}\cdot e^{-\frac{d^2}{2n}}, \tag{A.8}$$

under the imposed conditions,

$$\frac{e^{-\frac{d^2}{2n}}}{\sqrt{n}}\le\frac{\sqrt{2\pi}d}{n}\cdot\left(1+\frac{n}{d^2-n}\right)\cdot P_G\left(\frac{d}{\sqrt{n}}\right)<\frac{6.5d}{n}\cdot P_G\left(\frac{d}{\sqrt{n}}\right). \tag{A.9}$$

We finally have

$$\begin{aligned}2^{-n}\cdot S(n,d)&\le P_G\left(\frac{d}{\sqrt{n}}\right)\cdot\left(1+\frac{6.55d}{n}+e^{-\frac{3n}{16}+\frac{d^2}{2n}}\right)\\&<P_G\left(\frac{d}{\sqrt{n}}\right)\cdot\left(1+\frac{6.6d}{n}\right),\end{aligned} \tag{A.10}$$

where we have bounded $3/4<d/20$, $d^2/(2n)<n/8$, and $e^{-\frac{n}{16}}<d/(25n)$.

Now, let us prove (2.4). Starting from the lower bound in Lemma 2.1,

$$\begin{aligned}&2^{-n}\cdot S(n,d)\\&\quad\ge\sum_{k=\frac{d}{2}}^{(n^3/2)^{\frac14}}\left(1-\frac{3k^4}{2n^3}-\frac{1}{2n}\right)\cdot\sqrt{\frac{2}{\pi n}}\cdot e^{-2k^2/n}\\&\quad\ge\left(1-\frac{1}{2n}\right)\left[P_G\left(\frac{d}{\sqrt{n}}\right)-P_G\left(\left(\frac{n}{8}\right)^{\frac14}\right)\right]-\sqrt{\frac{2}{\pi n}}\cdot\sum_{k=\frac{d}{2}}^{(n^3/2)^{\frac14}}\frac{3k^4}{2n^3}\cdot e^{-2k^2/n}.\end{aligned}$$

To complete the proof, let us provide an upper bound for

$$S_3(n,d)=\sqrt{\frac{2}{\pi n}}\cdot\sum_{k=\frac{d}{2}}^{(n^3/2)^{\frac14}}\frac{3k^4}{2n^3}\cdot e^{-2k^2/n}.$$

Note that the maximum of $k^4e^{-2k^2/n}$ is reached for $k^2=n$. If $d>2\sqrt{n}$, then the summands in $S_3(n,d)$ are monotonously decreasing and the sum can be bounded from above by an integral as follows:

$$\sqrt{\frac{2}{\pi n}}\cdot\sum_{k=\frac{d}{2}}^{(n^3/2)^{\frac14}}\frac{3k^4}{2n^3}\cdot e^{-2k^2/n}<\frac{3}{8n}\sqrt{\frac{1}{\pi}}\int_{\frac{d-1}{\sqrt{2n}}}^{\infty}x^4e^{-x^2}dx.$$

On the other hand, if $\sqrt{\ln\ln n}<d/\sqrt{n}\le 2$ (which can happen only when $\ln\ln n<2$, i.e., for $n<1619$), then the summands increase for $d/2<k\le\sqrt{n}$ and decrease thereafter. The biggest summand is $\le 4e^{-2}\le 4e^{-(d-1)^2/2}$.

Clearly, $S_3(n,d)$ can be bounded from above in both cases as

$$\sqrt{\frac{2}{\pi n}}\cdot\sum_{k=\frac{d}{2}}^{(n^3/2)^{\frac{1}{4}}}\frac{3k^4}{2n^3}\cdot e^{-2k^2/n}\le\frac{3}{8n}\sqrt{\frac{2}{\pi n}}\cdot 4\cdot e^{-\frac{(d-1)^2}{2n}}+\frac{3}{8n}\sqrt{\frac{1}{\pi}}\int_{\frac{d-1}{\sqrt{2n}}}^{\infty}x^4e^{-x^2}dx$$

$$=\frac{3}{32n}\left[\frac{d-1}{\sqrt{2\pi n}}\left(\frac{(d-1)^2}{n}+3+\frac{32}{d-1}\right)e^{-\frac{(d-1)^2}{2n}}\right.$$

$$\left.+\,3P_G\left(\frac{d-1}{\sqrt{n}}\right)\right].$$

Analogously to (A.9), we have

$$\frac{e^{-\frac{(d-1)^2}{2n}}}{\sqrt{n}}<\frac{8.26(d-1)}{n}\cdot P_G\left(\frac{d-1}{\sqrt{n}}\right)<\frac{8.26d}{n}\cdot P_G\left(\frac{d-1}{\sqrt{n}}\right). \tag{A.11}$$

Lumping the contributions

$$\frac{(d-1)^2}{n}+3+\frac{32}{d-1}<4.55\frac{(d-1)^2}{n}<4.55\frac{d^2}{n}, \tag{A.12}$$

we get

$$\sqrt{\frac{2}{\pi n}}\cdot\sum_{k=\frac{d}{2}}^{(n^3/2)^{\frac{1}{4}}}\frac{3k^4}{2n^3}\cdot e^{-2k^2/n}<\frac{3}{8n}\sqrt{\frac{1}{\pi}}\left[\frac{37.6d^4}{4\sqrt{2}n^2}+\frac{3\sqrt{\pi}}{4}\right]P_G\left(\frac{d-1}{\sqrt{n}}\right)$$

$$<\frac{45d^4}{16\sqrt{\pi}n^3}P_G\left(\frac{d-1}{\sqrt{n}}\right)<\frac{5d^4}{n^3}P_G\left(\frac{d}{\sqrt{n}}\right),$$

where in the last inequality we used

$$P_G\left(\frac{d-1}{\sqrt{n}}\right)\le e^{(2d-1)/2n}\cdot\frac{d-1}{d}\cdot\frac{d^2/n}{d^2/n-1}\cdot P_G\left(\frac{d}{\sqrt{n}}\right)<3.04\cdot P_G\left(\frac{d}{\sqrt{n}}\right).$$

Finally, noting that

$$P_G\left(\left(\frac{n}{8}\right)^{\frac{1}{4}}\right)\le\sqrt{\frac{\sqrt{8}}{\pi}}\cdot n^{-1/4}\cdot e^{-\sqrt{n/32}}<e^{-\sqrt{n/32}},$$

we have

$$2^{-n}\cdot S(n,d)\ge\left(1-\frac{1}{2n}-\frac{5d^4}{n^3}\right)\cdot P_G\left(\frac{d}{\sqrt{n}}\right)-e^{-\sqrt{n/32}}. \tag{A.13}$$

∎

A.3. Probability of intersection. Additional elaboration. First, note that

$$\sum_{b=-(a-\frac{d}{2})}^{a-\frac{d}{2}}e^{-4b^2/n}\le\int_{-(a-\frac{d}{2})}^{a-\frac{d}{2}}e^{-4z^2/n}dz+e^{-(2a-d)^2/n} \tag{A.14}$$

$$=\frac{\sqrt{n}}{2}\cdot\int_{-\frac{2a-d}{\sqrt{n}}}^{\frac{2a-d}{\sqrt{n}}}e^{-x^2}dx+e^{-4(a-d/2)^2/n}.$$

In (A.14) we used the facts that $e^{-4b^2/n}$ is a symmetric convex function, and we are summing symmetrically around zero.

By now we have (using $d < n/4$)

$$P(d) \le \left(1 + o(n^{-1/4})\right) \cdot \frac{2}{\pi\sqrt{n}} \cdot \sum_{a=\frac{d}{2}}^{\frac{n}{4}} e^{-4a^2/n} \cdot \int_{-\frac{2a-d}{\sqrt{n}}}^{\frac{2a-d}{\sqrt{n}}} e^{-x^2} dx$$

$$+ \left(1 + o(n^{-1/4})\right) \cdot \frac{4}{\pi n} \sum_{a=\frac{d}{2}}^{\frac{n}{4}} e^{-4(a^2+(a-d/2)^2)/n} + e^{-\frac{\sqrt{n}}{20 \ln \ln n}}.$$

Using

$$\sum_{a=\frac{d}{2}}^{\frac{n}{4}} e^{-4a^2/n} < \sum_{a=\frac{d}{2}}^{\infty} e^{-4a^2/n} \le \frac{\sqrt{n}}{2} \int_{\frac{d}{\sqrt{n}}}^{\infty} e^{-x^2} dx + e^{-d^2/n}$$

and

$$\int_{-\frac{2a-d}{\sqrt{n}}}^{\frac{2a-d}{\sqrt{n}}} e^{-z^2} dz \le \int_{-\infty}^{\infty} e^{-z^2} dz = \sqrt{\pi}, \tag{A.15}$$

we have

$$\sum_{a=\frac{d}{2}}^{\frac{n}{4}} e^{-4a^2/n} \int_{-\frac{2a-d}{\sqrt{n}}}^{\frac{2a-d}{\sqrt{n}}} e^{-z^2} dz \le \frac{\sqrt{n}}{2} \int_{\frac{d}{\sqrt{n}}}^{\infty} e^{-x^2} dx \int_{-\frac{2x-d}{\sqrt{n}}}^{\frac{2x-d}{\sqrt{n}}} e^{-z^2} dz + \sqrt{\pi} e^{-d^2/n}. \tag{A.16}$$

Summing up, we have shown that

$$P(d) \le \left(1 + o(n^{-1/4})\right) \cdot \frac{1}{\pi} \int_{\frac{d}{\sqrt{n}}}^{\infty} e^{-x^2} dx \int_{-\frac{2x-d}{\sqrt{n}}}^{\frac{2x-d}{\sqrt{n}}} e^{-z^2} dz$$

$$+ \left(1 + o(n^{-1/4})\right) \cdot \frac{2}{\sqrt{\pi n}} \cdot \left(e^{-d^2/n} + \frac{2}{\sqrt{\pi n}} \sum_{a=\frac{d}{2}}^{\frac{n}{4}} e^{-4(a^2+(a-d/2)^2)/n} \right) + e^{-\frac{\sqrt{n}}{20 \ln \ln n}}$$

$$= \left(1 + o(n^{-1/4})\right) \cdot \frac{1}{\pi} \int_{\frac{d}{\sqrt{n}}}^{\infty} e^{-x^2} dx \int_{-\frac{2x-d}{\sqrt{n}}}^{\frac{2x-d}{\sqrt{n}}} e^{-z^2} dz.$$

REFERENCES

[1] N. Alon and J. Spencer, *The Probabilistic Method*, 2nd ed., John Wiley and Sons, New York, 2000.

[2] E. R. Berlekamp and L. R. Welch, *Weight distributions of the cosets of the* $(32, 6)$ *Reed-Muller code*, IEEE Trans. Inform. Theory, 18 (1972), pp. 203–207.

[3] C. Carlet, *On cryptographic complexity of Boolean functions*, in Proceedings of the Sixth Conference on Finite Fields with Applications to Coding Theory, Cryptography, and Related Areas, G. L. Mullen, H. Stichtenoth, and H. Tapia-Recillas, eds., Springer-Verlag, Berlin, 2002, pp. 53–69.

[4] C. Carlet, *On the degree, nonlinearity, algebraic thickness, and non-normality of Boolean functions, with developments on symmetric functions*, IEEE Trans. Inform. Theory, 50 (2004), pp. 2178–2185.
[5] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering Codes*, North–Holland, Amsterdam, 1997.
[6] G. Halász, *On the result of Salem and Zygmund concerning random polynomials*, Studia Sci. Math. Hungar., 8 (1973), pp. 369–377.
[7] M. Hall, Jr., *Combinatorial Theory*, 2nd ed., John Wiley and Sons, New York, 1986.
[8] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North–Holland, Amsterdam, 1977.
[9] J. A. Maiorana, *A classification of the cosets of the Reed-Muller code* $R(1,6)$, Math. Comp., 57 (1991), pp. 403–414.
[10] D. Olejár and M. Stanek, *On cryptographic properties of random Boolean functions*, J. UCS, 4 (1998), pp. 705–717.
[11] F. Rodier, *On the non-linearity of Boolean functions*, in Proceedings of the Workshop on Coding and Cryptography (WCC2003), INRIA, 2003, pp. 397–405.
[12] F. Rodier, *Sur la non-linéarité des fonctions booléennes*, Acta Arith., 115 (2004), pp. 1–22.
[13] F. Rodier, *Asymptotic nonlinearity of Boolean functions*, Des. Codes Cryptogr., 40 (2006), pp. 59–70.
[14] J. Spencer, *Six standard deviations suffice*, Trans. Amer. Math. Soc., 289 (1985), pp. 679–706.
[15] W. D. Wallis, A. P. Street, and J. Seberry Wallis, *Combinatorics: Room squares, sum-free sets, Hadamard matrics*, Lecture Notes in Math. 292, Springer-Verlag, Berlin, 1972.
[16] C.-K. Wu, *On distribution of Boolean functions with nonlinearity* $\leq 2^{n-2}$, Australas. J. Combin., 17 (1998), pp. 51–59.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 96–108

K_6-MINORS IN TRIANGULATIONS ON THE KLEIN BOTTLE*

KEN-ICHI KAWARABAYASHI†, RAIJI MUKAE‡, AND ATSUHIRO NAKAMOTO§

Abstract. In this paper, we shall characterize triangulations on the Klein bottle without K_6-minors. Our characterization implies that every 5-connected triangulation on the Klein bottle has a K_6-minor. The connectivity "5" is best possible in a sense that there is a 4-connected triangulation on the Klein bottle without K_6-minors.

Key words. triangulation, K_6-minor, Klein bottle

AMS subject classifications. 05C10, 05C83

DOI. 10.1137/070693540

1. Introduction.

Our motivation comes from the following result by Wagner, and a wide open question concerning a characterization of graphs without K_6-minors.

THEOREM 1 (see Wagner [16]). *A graph G has no K_5-minors if and only if G can be obtained from planar graphs and subgraphs of V_8 by means of clique-sums of order at most three.*

To understand this result, we need some notation. A graph H is said to be a *minor* of a graph K if H can be obtained from a subgraph of K by contracting edges. In this case, we say that K has an *H-minor*.

Let G_1 and G_2 be graphs with disjoint vertex sets, let $k \geq 1$ be an integer, and for $i = 1, 2$, let $X_i \subseteq V(G_i)$ be a k-clique in G_i, i.e., a set of k mutually adjacent vertices. For $i = 1, 2$, let G_i' be obtained from G_i by deleting a (possibly empty) set of edges with both ends in X_i. Let G be the graph obtained from G_1' and G_2' by identifying X_1 and X_2. Then we say that G is a *clique-sum* of *order* k, or simply a *k-sum* of G_1 and G_2. Let V_8 be the graph obtained from the 8-cycle C_8 by joining each pair of diagonally opposite vertices by an edge, which is sometimes called a *Möbius ladder*.

Theorem 1 implies that the four color theorem is equivalent to the statement that every graph without K_5-minors can be colored with four colors (Wagner's equivalence theorem). This result prompted Hadwiger [6] to make his famous conjecture: every graph without K_k-minor is $(k-1)$-colorable. This conjecture is considered by many as one of the deepest open problems in graph theory. To attack this conjecture, we would like to know more about the structure of graphs with no K_k-minors.

The obvious choice would be the next case: what kind of graphs do not contain a K_6-minor? We would like to make an attempt on this problem, but unfortunately, this question is wide open, and even hopeless right now.

Robertson, Seymour, and Thomas [12] proved the following result when dealing with Hadwiger's conjecture for K_6-minor-free case. Let us recall that a graph G is an *apex graph* if it has a vertex v such that $G - v$ is planar.

*Received by the editors June 2, 2007; accepted for publication (in revised form) June 19, 2008; published electronically October 24, 2008.
http://www.siam.org/journals/sidma/23-1/69354.html

†National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan (k_keniti@nii.ac.jp).

‡Department of Information Media and Environment Sciences, Graduate School of Environment and Information Sciences, Yokohama National University, 79-7 Tokiwadai, Hodogaya-ku, Yokohama 240-8501, Japan (d07tc019@ynu.ac.jp).

§Department of Mathematics, Faculty of Education and Human Sciences, Yokohama National University, 79-2 Tokiwadai, Hodogaya-ku, Yokohama 240-8501, Japan (nakamoto@edhs.ynu.ac.jp).

THEOREM 2 (see Robertson, Seymour, and Thomas [12]). *Let G be a graph with no K_6-minor such that G is not 5-colorable, and subject to that, the number of vertices of G is as small as possible. Then G is an apex graph.*

This theorem implies that Hadwiger's conjecture for K_6-minor-free case is equivalent to the four color theorem. But, unfortunately, this theorem does not give any structural characterization for graphs with no K_6-minor. In fact, Jørgensen [7] made the following beautiful conjecture.

CONJECTURE 3 (see Jørgensen [7]). *Every 6-connected graph containing no K_6-minor is an apex graph.*

Mader [10] proved that the graph G mentioned in Theorem 2 is 6-connected. Hence the above conjecture implies Theorem 2. This conjecture is still open, but recently, DeVos et al. [3] proved the following remarkable result. (However, the proof is lengthy and complicated since it needs some deep results in graph minor theory.)

THEOREM 4 (see DeVos et al. [3]). *Jørgensen's conjecture is true for large graphs. More precisely, there exists a constant N such that every 6-connected graph with no K_6-minor and with at least N vertices is apex.*

One may ask the following: what about 5-connected graphs with no K_6-minors? As far as we know, there are six families of graphs that do not contain K_6-minors. These are planar graphs, apex graphs, double cross graphs, planar graphs plus a triangle, graphs with hamburger structure, and graphs with hose structure. For *double cross graphs* and the *hose structure*, see Figure 1, in which shaded "blobs" represent planar graphs embedded in a disk with specified vertices on the boundary. For consecutive "blobs" in the hose structure, the five vertices are identified, not necessarily in order as suggested by their closeness in the figure, but the three white vertices are identified with white and the two black with black in the neighboring "blob." Graphs with *hamburger structure* are obtained from three 5-connected planar graphs G_i $(i = 1, 2, 3)$, each of which has a specified vertex w_i of degree 5. Let $v_{i1}, \ldots, v_{i5}$ be the neighbors of w_i in the clockwise order around w_i. To get a graph with hamburger structure, take $G_1 - w_1$, $G_2 - w_2$, and $G_3 - w_3$ and identify for $j = 1, \ldots, 5$ their vertices v_{1j}, v_{2j}, v_{3j}. These examples give rise to infinitely many 5-connected graphs without K_6-minors and with a different structure. (For apex graphs, double cross graphs, and planar graphs plus a triangle, we can easily prove that all of them contain no K_6-minors, but for the hamburger structure and hose structure, the proof for no K_6-minor is not so easy.)

At this moment it seems hopeless to characterize 5-connected graphs with no K_6-minor, even for large graphs. This gives us an impression that a complete characterization of graphs without K_6-minors is very hard, even hopeless, since we definitely need to figure out which 5-connected graphs do not contain K_6-minors.

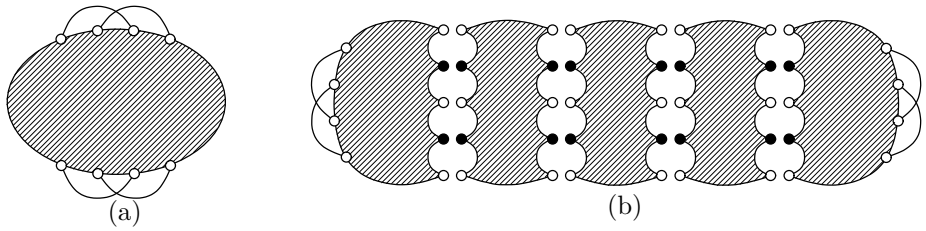


FIG. 1. (a) *Double cross and* (b) *hose structure graphs.*

Thus we set a more modest goal in this paper: we restrict ourselves to consider graphs on a fixed surface. Actually, our main interest in this paper is a *triangulation*

on the Klein bottle, that is, a simple graph embedded in the surface such that each face is triangular. Let us see our motivation concerning this family of graphs. As we pointed out, a double cross graph is one of the obstructions for a 5-connected graph without K_6-minors. But it is easy to see that this graph is embeddable in the Klein bottle, but cannot be embedded into a projective plane nor a torus. So it might be interesting to consider 5-connected graphs on the Klein bottle. But this needs a great deal of case analysis, and we are not sure yet whether or not this problem is feasible, i.e., is it really much easier than the general 5-connected case? In [5], Fijavź and Mohar proved that every 5-connected graph on a projective plane with representativity at least 3 has a K_6-minor, but the proof still needs a great deal of case analysis, and some of the deep results in graph minor papers. Their result does not seem to be enough to give a complete characterization of projective planar graphs without K_6-minors.

So, it seems that even the torus and Klein bottle cases are hard. But if we restrict our attention to triangulations, then the situation is much different. In fact, the results in [11] and the result in this paper give a complete characterization of triangulations on the projective plane or the torus or the Klein bottle that do not contain K_6-minors. Let us see these results.

For the projective plane and the torus, the following have been proved in [11]. A *quadrangulation* on a surface is a simple graph with each face quadrilateral. An *H-quadrangulation* is a quadrangulation isomorphic to H as a graph.

THEOREM 5 (see Mukae and Nakamoto [11]). *A triangulation G on the projective plane has a K_6-minor if and only if G has no K_4-quadrangulation as a subgraph.*

THEOREM 6 (see Mukae and Nakamoto [11]). *A triangulation G on the torus has a K_6-minor if and only if G has no K_5-quadrangulation as a subgraph.*

Figure 2 shows a K_4-quadrangulation on the projective plane in the left-hand side, and a K_5-quadrangulation on the torus in the right-hand side. (In Figure 2, in order to obtain the projective plane and the torus, we identify any pair of antipodal points of the hexagon in the left-hand side, and identify the two horizontal segments, and the two vertical segments, in the right-hand side respectively.)

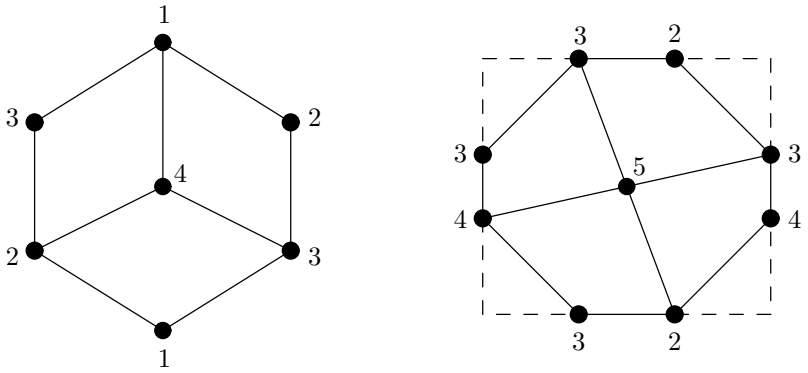


FIG. 2. *K_4- and K_5-quadrangulations.*

In this paper, we shall characterize triangulations on the Klein bottle without K_6-minors and prove the following theorem corresponding to Theorems 5 and 6.

A *Möbius triangulation* (G, C) is a triangulation G on the Möbius band with boundary cycle C. Let Q be a 2-connected graph on the sphere, and let F_1, F_2 be two distinct faces of Q, where C_i is the boundary cycle of F_i, for $i = 1, 2$. Suppose that

each face except F_1, F_2 is bounded by a 3-cycle. Let G be the graph obtained from Q by removing the interior of F_1 and F_2. We say that $G = (G, C_1, C_2)$ is an *annulus triangulation*, where each C_i is called the *boundary* (or *boundary cycle* of G). A cycle C of G is said to be *essential* if C is homotopic to C_1 and C_2 in the annulus.

THEOREM 7. *A triangulation G on the Klein bottle has no K_6-minor if and only if G has two 4-cycles C_1 and C_2 separating G into two Möbius triangulations (M_i, C_i) for $i = 1, 2$, and one annulus triangulation (A, C_1, C_2) such that*

(i) *the four vertices of C_i induce K_4 in (M_i, C_i) for $i = 1, 2$, and*

(ii) *(A, C_1, C_2) satisfies one of the following:*

(a) *(A, C_1, C_2) has an essential 3-cycle, or*

(b) *(A, C_1, C_2) has m essential 4-cycles $D_1, \ldots, D_m$ for some $m \geq 2$ lying on the annulus in this order such that $C_1 = D_1, C_2 = D_m$, and for each i, $V(D_i) \cap V(D_{i+1}) \neq \emptyset$.*

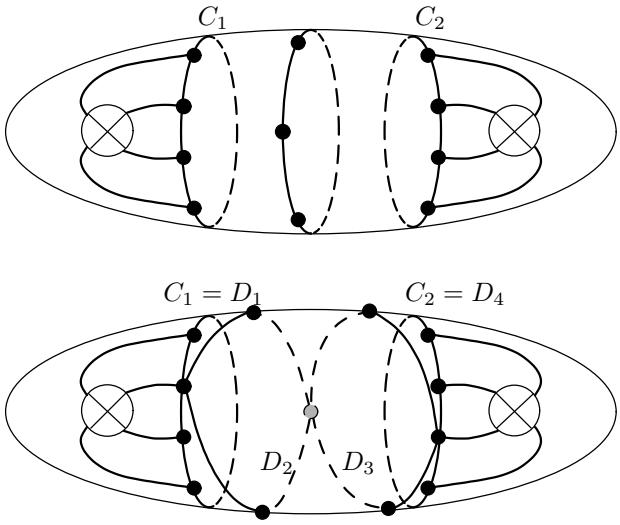


FIG. 3. *Structures of triangulations on the Klein bottle with no K_6-minor.*

Note that in Theorem 7, if G is 4-connected, then (ii)(b) must happen, since the 3-cycle in (ii)(a) separates G. (See Figure 3 which shows the structure of triangulations on the Klein bottle with no K_6-minor. The top shows a triangulation on the Klein bottle with an essential separating 3-cycle corresponding to (ii)(a) in Theorem 7, and the bottom is one with four essential 4-cycles corresponding (ii)(b).)

The following is an immediate consequence from Theorem 7, since each triangulation on the Klein bottle with no K_6-minor has a separating 3- or 4-cycle.

COROLLARY 8. *Every 5-connected triangulation on the Klein bottle has a K_6-minor.*

The connectivity "5" is best possible, since there is a 4-connected triangulation on the Klein bottle without K_6-minors, as we pointed out above. For the projective plane and the torus, Theorems 5 and 6 imply the same fact as Corollary 8. In view of known 5-connected graphs without K_6-minors, it is perhaps true that every 5-connected triangulation on any nonspherical surface has a K_6-minor.

2. Irreducible triangulations on the Klein bottle. Let us first consider a topology of the Klein bottle, which admits three different types of simple closed

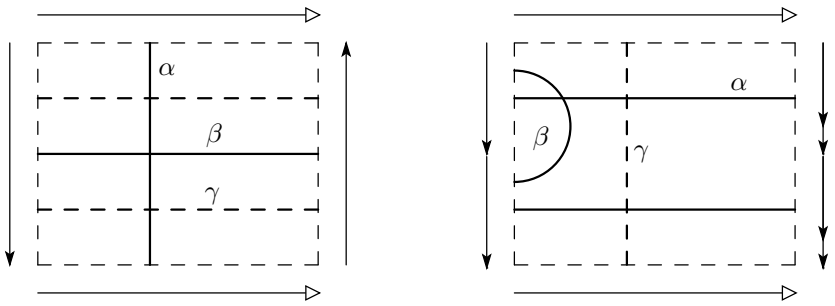


FIG. 4. *Klein bottle with a meridian α, a longitude β, and an equator γ.*

curves. Let $\mathbb{N}_k$ denote the nonorientable surface of genus k throughout this paper. Then $\mathbb{N}_1$ and $\mathbb{N}_2$ stand for the projective plane and the Klein bottle, respectively. A simple closed curve l on a nonspherical surface F^2 is said to be *essential* if l does not bound a 2-cell on F^2. We say that l is 1-*sided* if a tabular neighborhood of l is homeomorphic to a Möbius band, and 2-*sided* otherwise. See Figure 4, which shows two developments of $\mathbb{N}_2$. (We identify the top and bottom of the rectangle naturally to get an annulus, and there are two ways to get $\mathbb{N}_2$ from the annulus. One is to identify the two boundary components incoherently as in the left-hand side, and the other is to identify each pair of antipodal points of each boundary component as in the right-hand side. In particular, the expression of $\mathbb{N}_2$ in Figure 3 corresponds to the right-hand side of Figure 4.) Let α, β, γ be three essential simple closed curves on $\mathbb{N}_2$ as in Figure 4, where each of α, β, and γ in both figures stands for the same closed curve on $\mathbb{N}_2$. Observe that α is a 2-sided simple closed curve cutting $\mathbb{N}_2$ into an annulus, β is a 1-sided one cutting $\mathbb{N}_2$ into a Möbius band, and β is a 2-sided one separating $\mathbb{N}_2$ into two Möbius bands. We say that γ is an *equator*, and a cycle of a graph on $\mathbb{N}_2$ homotopic to γ is an *equator cycle*.

Let G be a triangulation, and let e be an edge of G. *Contraction* of e (or *contracting* e) in G is to remove e, identify the two ends of e and replace two pairs of multiple edges by single edges respectively. We say that e is *contractible* if the graph obtained from G by contracting e is simple. Moreover, we say that G is *contractible* to a triangulation H if G can be transformed into H by a sequence of contracting edges. For a graph G on a surface and a vertex v, the *link* of v is the boundary closed walk of the union of all faces incident to v in G. A cycle C of a graph G on a surface is said to be *essential* if C does not bound a 2-cell on the surface. For a graph G and a subset S of $V(G)$, let $[S]$ denote the subgraph of G induced by S. For a path or cycle C in a graph G, a *chord* of C is an edge xy such that $x, y \in V(C)$ and $xy \notin E(C)$.

We say that a triangulation G is *irreducible* if G has no contractible edge. The complete lists of irreducible triangulations on the sphere, the projective plane and the torus have already been determined in [13], [1], and [8], respectively. For the Klein bottle, Lawrenceko and Negami [9] and Sulanke [14] determined the complete list of irreducible triangulations, in which 25 triangulations, denoted by $Kh1, \ldots, Kh25$, are 4-connected and the other four have equator 3-cycles.

LEMMA 9. *All 4-connected irreducible triangulations on $\mathbb{N}_2$ except $Kh25$ have K_6-minors.*

Proof. We have checked that $Kh3$, $Kh5$, $Kh6$, $Kh7$, $Kh9$, $Kh13$, $Kh16$, $Kh17$ include K_6 as a subgraph, as shown in Figure 5. On the other hand, each of $Kh1$, $Kh2$, $Kh4$, $Kh8$, $Kh10$, $Kh11$, $Kh12$, $Kh14$, $Kh15$, $Kh18$, $Kh19$, $Kh20$, $Kh21$, $Kh22$,

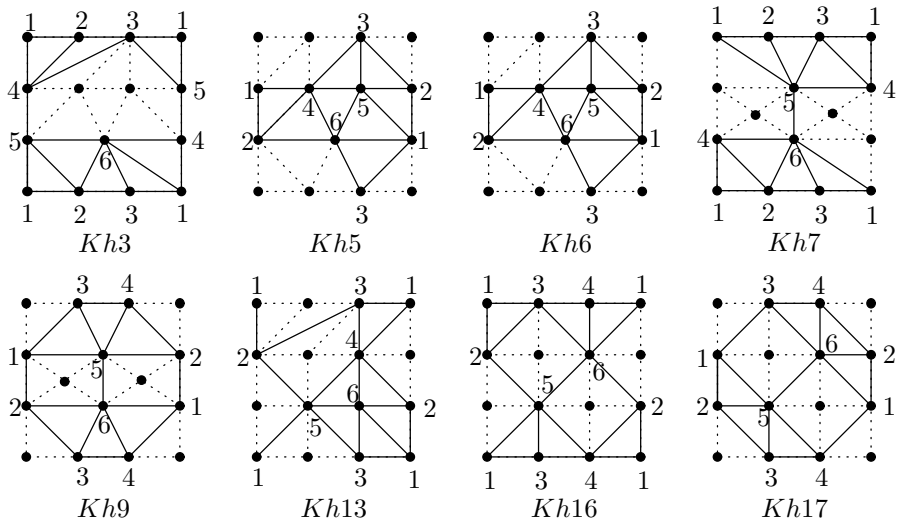


FIG. 5. *K_6-subgraphs in irreducible triangulations on the Klein bottle.*

$Kh23$, $Kh24$ has a K_6-minor as shown in Figure 6, in which vertices surrounded by a single polygon corresponds to a single vertex of K_6 after the contractions. □

The triangulation $Kh25$ is shown in Figure 7 in which we identify the sides as in the left of Figure 4. We can see that the following holds for $Kh25$.

LEMMA 10. *The irreducible triangulation $Kh25$ in Figure 7 has two equator 4-cycles $C_1 = abcd$ and $C_2 = abef$ such that for $i = 1, 2$, the four vertices of C_i induce K_4 in the Möbius triangulation cut off from G by C_i.*

3. Lemmas. We comprise several lemmas for proving Theorem 7. The first one is the most famous theorem in graph theory, called "Kuratowski–Wagner's theorem."

LEMMA 11 (see [15]). *A graph G is planar if and only if G contains neither a K_5-minor nor a $K_{3,3}$-minor.*

Let us mention a fundamental result due to Wagner [16].

LEMMA 12 (see [16]). *Suppose that a graph G is obtained from two graphs H_1 and H_2 by a k-sum for some $k \leq n-2$. Then G has a K_n-minor if and only if one of H_1 and H_2 has a K_n-minor.*

The following lemma immediately follows from Theorem 5 and Lemma 12.

LEMMA 13. *Let G be a triangulation on $\mathbb{N}_2$ with an equator 3-cycle C, and let G_1 and G_2 be the two triangulations on $\mathbb{N}_1$ obtained from G by cutting along C and capping off by 2-cells. Then G has no K_6-minor if and only if both of G_1 and G_2 have K_4-quadrangulations as subgraphs.*

Let D be a plane graph with boundary cycle C and each inner face triangular, and let x, y be distinct vertices of C. An $x-y$ path P is said to be *internal* if P intersects C only at its endvertices x and y.

LEMMA 14 (see [2]). *Let D be a plane graph with boundary cycle C and each inner face triangular, and let x, y be distinct vertices of C with $xy \notin E(C)$. Then D has an internal $x-y$ path if and only if D has no chord pq for some $p, q \in V(C) - \{x, y\}$ such that x and y are contained in different components of $C - \{p, q\}$.*

Let G be a 2-connected plane graph with outer cycle C of length at least 4 such that each inner face is triangular. Suppose that four distinct vertices v_1, v_2, v_3, v_4,

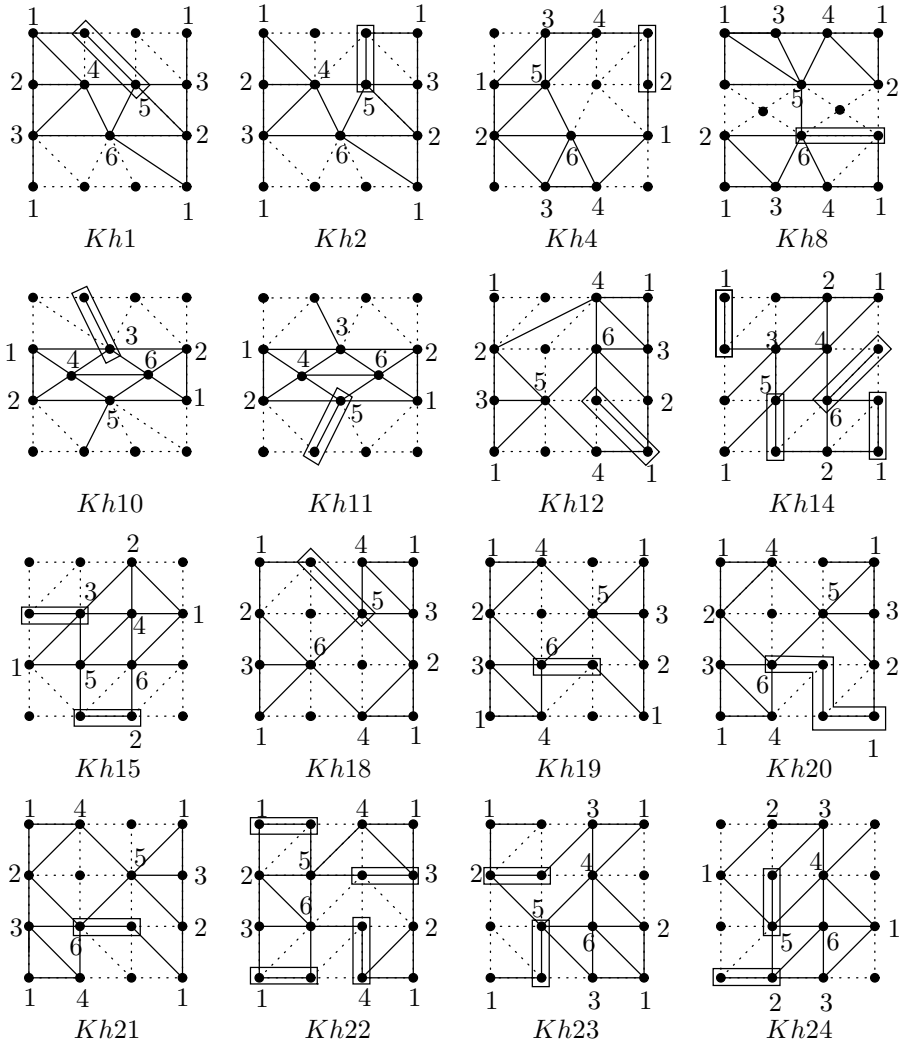


FIG. 6. *K_6-minors in irreducible triangulations on the Klein bottle.*

called *nodes*, lie on C in this order but v_i and v_{i+1} do not need to be consecutive in C, for each i. We call G a 4-*patch with nodes* $\{v_1, v_2, v_3, v_3\}$. The subpath in C between v_i and v_{i+1} (not containing v_{i+2}) is denoted by $[v_i, v_{i+1}]$. A $v_i - v_{i+2}$ path in G avoiding v_{i+1} and v_{i+3} is called a *diagonal* or a $v_i - v_{i+2}$ *diagonal*. (Note that a $v_i - v_{i+2}$ diagonal might not be internal $v_i - v_{i+2}$ path in the 4-patch G.)

The following lemma immediately follows from Lemma 14.

LEMMA 15. *Let G be a 4-patch with nodes $\{v_1v_2v_3v_4\}$. Then, unless $v_1v_3 \in E(G)$, G has a $v_2 - v_4$ diagonal.*

Let G be a 4-patch with nodes $\{v_1, v_2, v_3, v_4\}$, and let C be the outer cycle of G. Suppose that $v_1v_2, v_3v_4 \in E(C)$ and $v_1v_3 \notin E(G)$. Let us define a special $v_2 - v_4$ diagonal in G. Let $R = p_1 \cdots p_m$ be the path in G consisting of the neighbors of v_1, where $p_1 = v_2$, and p_m is the first vertex lying on $[v_1, v_4]$. Let p_a be the last vertex contained in $[v_2, v_3]$. Let $P = p_1 \cdots p_a$, and let Q be the subpath of $[v_1, v_4]$ joining

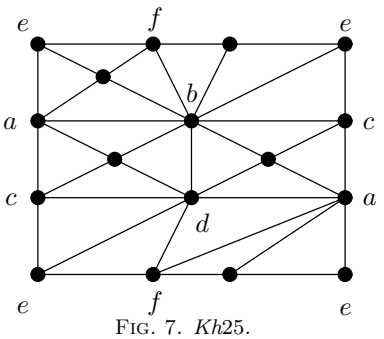


FIG. 7. *Kh*25.

p_m and v_4. Since $v_1v_3 \notin E(G)$, $R \cup Q$ forms a $v_2 - v_4$ diagonal, which is called the $v_2 - v_4$ diagonal *closest* to v_1. We say that P and Q are called the *initial* and the *terminal* segments of the diagonal.

The following is an important lemma.

LEMMA 16. *Let $A = (A, D, D')$ be an annulus triangulation such that D and D' are disjoint 4-cycles. Suppose that A has neither an essential 3-cycle nor an essential 4-cycle except D and D'. Then A has disjoint four paths P_i joining $V(D)$ and $V(D')$ for $i = 1, 2, 3, 4$. Moreover, if we let R_i be the 4-patch bounded by P_i and P_{i+1} for $i = 1, 2, 3, 4$, then each R_i has a diagonal D_i such that for any disjoint $i, j \in \{1, 2, 3, 4\}$, $V(D_i) \cap V(D_j) = \emptyset$.*

Proof. Observe that A has at least eight vertices, and that the lemma clearly holds when A has exactly eight vertices. So we suppose A has at least nine vertices.

Since (A, D, D') has no essential 3-cycle, (A, D, D') has disjoint four paths P_i joining $V(D)$ and $V(D')$, for $i = 1, 2, 3, 4$, by Menger's theorem. Suppose that $D = v_1v_2v_3v_4$ and $D' = v'_1v'_2v'_3v'_4$, and that P_i joins $v_i \in V(D)$ and $v'_i \in V(D')$, for $i = 1, 2, 3, 4$. Let R_i be the plane subgraph of (A, D, D') bounded by $P_i, v_iv_{i+1}, P_{i+1}, v'_{i+1}v'_i$, for $i = 1, 2, 3, 4$. Then each R_i is a 4-patch with nodes $\{v_i, v_{i+1}, v'_{i+1}, v'_i\}$. We shall prove that R_1, R_2, R_3, R_4 have pairwise disjoint diagonals in A if A has no essential 4-cycle except D and D'. Such an essential 4-cycle is said to be *bad* in this proof.

Suppose the lemma does not hold, and let A be the smallest counterexample of the lemma. That is, A is an annulus triangulation with the fewest number of vertices which has no bad 4-cycle, but R_1, R_2, R_3, R_4 do not have disjoint diagonals.

CLAIM 1. *$v_iv'_{i+1}, v'_iv_{i+1} \notin E(R_i)$ for $i = 1, 2, 3, 4$.*

Proof. For contradictions, we may suppose that R_1 has an edge $v_1v'_2$ without loss of generality. Observe that R_i has no edge v'_iv_{i+1} for $i = 2, 3, 4$ (otherwise, we would find a bad 4-cycle through $v_1v'_2$ and v'_iv_{i+1}, a contradiction). Hence each R_i has a $v_i - v'_{i+1}$ diagonal, by Lemma 15. Let $D_1 = v_1v'_2$. Take a $v'_3 - v_2$ diagonal D_2 in R_2 closest to v'_2, and a $v_4 - v'_1$ diagonal D_4 in R_4 closest to v_1. Replace P_3 and P_4 so that the initial segment of D_2, denoted by P'_3, lies on P_3, and the initial segment of D_4, denoted by P'_4, lies on P_4. (See Figure 8.) Since A has no bad 4-cycle, R_3 has no edge joining P'_3 and P'_4. Hence, by Lemma 14, R_3 has a $v_3 - v'_4$ diagonal avoiding the vertices of P'_3 and P'_4. Therefore, D_1, D_2, D_3, and D_4 are required disjoint paths, contrary to the assumption of A. □

If each P_i has length one, then each R_i admits both $v_i - v'_{i+1}$ and $v'_i - v_{i+1}$ internal diagonals, since R_i has no edge $v_iv'_{i+1}$ and v'_iv_{i+1}, by Claim 1 and Lemma 15. Hence A has desired four disjoint paths, contrary to the assumption of A. Therefore, for some i, P_i (say $i = 1$) has length of at least 2.

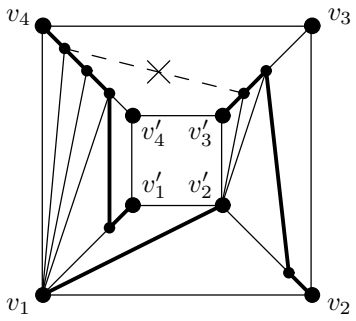


FIG. 8. *Case when* $v_1v'_2 \in E(R_1)$.

Let x be the neighbor of v_1 in P_1, where $x \neq v'_1$. Now contract the edge v_1x and let A' be the resulting annulus triangulation with disjoint boundary cycles. If A' has no bad 4-cycle, then A' has desired four paths, by the minimality of A. Clearly, the preimage of the four paths in A are required four disjoint paths, contrary to the assumption of A. Hence $[v_1x]$ is contained in bad 4-cycles in A', where $[v_1x]$ is the vertex of A' obtained from v_1x by its contraction in A. Let C' be the bad 4-cycle of A' through $[v_1x]$ bounding a fewest number of faces with D'. Let C be the 5-cycle through v_1 and x which is the preimage of C' in A. Moreover, C intersects P_i exactly once for $i = 2, 3, 4$. Hence putting $\{x_i\} = V(P_i) \cup V(C)$ for $i = 2, 3, 4$, we let $C = v_1x_2x_3x_4x$. (It may happen that $C = v_1xx_2x_3x_4$, but in this case we can relabel to get $C = v_1x_2x_3x_4x$.)

Suppose that C' does not intersect D'. Then the annulus triangulation A'' in A' bounded by C' and D' has no bad cycle. (Otherwise, if A'' had a bad 4-cycle C'', then either C'' would be a bad 4-cycle in A, or the preimage of C'' in A is a 5-cycle through v_1 and x which is closer to D' than C. The former contradicts the assumption of A, and the latter contradicts the assumption of C.) Hence A'' has required four disjoint paths, by the minimality of A. This means that A' contains the required four paths, and hence their preimages contradict the assumption on A.

Hence C' intersects D'. So, avoiding a bad 4-cycle and using Claim 1, we shall classify possible structure of A. We first observe $x_4 \neq v_4$. (For otherwise, A would have a bad 4-cycle $v_1x_2x_3v_4$, a contradiction.) We second have $x_3 \neq v_3$. (For, if $x_3 = v_3$, then we must have $x_2 = v_2$, since A has no bad 4-cycle. On the other hand, in this case, we must have $x_4 = v'_4$, since C' intersects D'. However, we must have $v_3v'_4 \in E(A)$, contrary to Claim 1.) Moreover, we third have $x_2 \neq v'_2$, since $v_1v'_2 \notin E(A)$ by Claim 1. So, if $x_2 = v_2$, then we must have (i), since $v_2v'_3 \notin E(A)$ and $C' \cap D' \neq \emptyset$. On the other hand, if x_2 is an inner vertex of P_2, then we have (ii), (iii), and (iv), as in the following (see Figure 9):

(i) $x_2 = v_2$, $x_3 \neq v_3, v'_3$, and $x_4 = v'_4$.
(ii) $x_2 \neq v_2, v'_2$, $x_3 \neq v_3, v'_3$, and $x_4 = v'_4$.
(iii) $x_2 \neq v_2, v'_2$, $x_3 = v'_3$, and $x_4 \neq v_4, v'_4$.
(iii) $x_2 \neq v_2, v'_2$, $x_3 = v'_3$, and $x_4 = v'_4$.

We first consider the case (i). Rechoosing P_3 and P_2, we may suppose that the vertices of the subpath, denoted by P'_3, of P_3 joining x_3 and v'_3 are adjacent to v'_4, and that the vertices of P_2 are adjacent to vertices of P'_3. Similarly, the vertices of the subpath, denoted by P'_1, of P_1 joining x and v'_1 are adjacent to v'_4. Then, no edge joins P'_1 and P_2 (except $v'_1v'_2$) since A has no bad 4-cycle. Therefore, by Lemma 14,

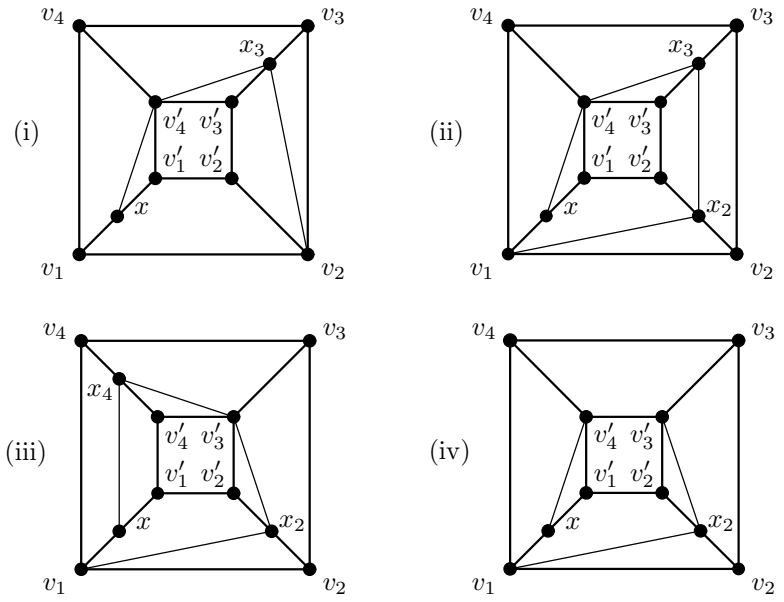


FIG. 9. 5-*cycle* $C = v_1x_2x_3x_4x$ *in* A.

the 4-patch with nodes $\{v_1, v_2, v'_2, v'_1\}$ admits an internal $v_1 - v'_2$ diagonal, and we put it as D_1. Let D_2 be a $v_2 - v'_3$ diagonal in the 4-patch with nodes $\{v_2, v'_2, v'_3, x_3\}$, which actually exists, by Lemma 14, since $v'_2x_3 \notin E(R_2)$. (If $v'_2x_3 \in E(R_2)$, then A would have a bad 4-cycle $v'_1v'_2x_3v'_4$.) Let D_3 be the path starting at v_3, reaching x_3 along P_3, and ending at v'_4 through the edge $v_3v'_4$. Let D'_4 be an $x - v_4$ diagonal in the 4-patch with nodes $\{v_1, x, v'_4, v_4\}$, which actually exists, by Lemma 14, since $v_1v'_4 \in E(R_4)$ by Claim 1. Let $D_4 = D'_4 \cup P'_1$. Then D_1, D_2, D_3, D_4 are required four disjoint diagonals.

The remaining three cases can be dealt with in almost the same way, but we have to be careful in (iii) to take a $v_3 - x_4$ diagonal D'_3 in the 4-patch with nodes $\{v_3, v'_3, x_4, v_4\}$ and a $v_4 - x$ diagonal D'_4 in the 4-patch with nodes $\{v_4, x_4, x, v_1\}$ so that they are disjoint in A. We first observe that two diagonals D'_3 and D'_4 exist since $v'_3v_4, v_1x_4 \notin E(A)$ by the assumption for bad 4-cycles in A. Take D'_3 to be closest to v'_3, and D'_4 to be closest to v_1. If D'_3 and D'_4 shared a vertex, say v, then v would be a common neighbor of v'_3 and v_1, and the 4-cycle $v_1x_2v'_3v$ would be bad in A, a contradiction. Hence we can take D'_3 and D'_4 to be disjoint in A. We need the same consideration in (iv). ∎

4. Proof of Theorem 7. In this section, we shall prove Theorem 7. For a short notation, the m equator 4-cycles $D_1, \ldots, D_m$ in Theorem 7(ii)(b) with $V(D_i) \cap V(D_{i+1}) \neq \emptyset$ for each i is called an *equator* 4*-cycle system*, or an m*-equator* 4*-cycle system* if m is emphasized.

Proof of sufficiency. We shall prove that if the graph is as described in Theorem 7, then it has no K_6-minors. Let G be a triangulation on $\mathbb{N}_2$ with two equator 4-cycles C_1 and C_2 separating G into two Möbius triangulations (M_i, C_i) such that the four vertices of C_i induce K_4 in M_i, for $i = 1, 2$, and one annulus triangulation (A, C_1, C_2).

By Lemma 13, we may suppose that A has no equator 3-cycles. Then A contains an m-equator 4-cycle system $D_1, \ldots, D_m$ with $C_1 = D_1$ and $C_2 = D_m$ for some

$m \geq 2$. Let $C_1 = v_1v_2v_3v_4$. Since C_1 separates G into two Möbius triangulations and since $V(C_1)$ induces K_4 in G, each of the 4-cycles $v_1v_2v_4v_3$ and $v_2v_3v_1v_4$ bound a 2-cell on $\mathbb{N}_2$. Let H_1 be the subgraph of G induced by v_1, v_2, v_3, v_4 and the vertices in the interior of $v_1v_2v_4v_3$. Note that H_1 is the graph obtained from the plane graph with boundary $v_1v_2v_4v_3$ and two edges v_1v_4 and v_2v_3, and that H_1 has at least five vertices, since the plane graph with boundary $v_1v_2v_4v_3$ has no edge v_1v_3 and v_2v_4, by the simpleness of G. Hence H_1 has no K_6-minor, since the graph obtained from H_1 by removing one edge v_1v_3 is planar. Let H_1' be the subgraph of G defined similarly for the 4-cycle $v_2v_3v_1v_4$. Let H_2 and H_2' be the two subgraphs of G defined similarly for $C_2 = u_1u_2u_3u_4$. Then each of H_1', H_2, H_2' has no K_6-minor either, similarly to H_1.

Let $\tilde{A}$ be the graph obtained from A by adding four edges $v_1v_3, v_2v_4, u_1u_3, u_2u_4$. Then G is obtained from $\tilde{A}$ by 4-sums of H_1, H_1', H_2, H_2' applied repeatedly. Therefore, by Lemma 12, since each of H_1, H_1', H_2, H_2' has no K_6-minor, we have only to prove that $\tilde{A}$ has no K_6-minor. We prove it as follows.

We use induction on the number m of the 4-cycles in an m-equator 4-cycle system in A. Suppose $m = 2$, and then D_1 $(= C_1)$ and D_2 $(= C_2)$. Here it is useful to consider a planar embedding of A with four edges $v_1v_3, v_2v_4, v_1'v_3', v_2'v_4'$ added, which is a planar drawing of $\tilde{A}$ with two crossings of edges. Since D_1 and D_2 share a vertex, say $v_1 = u_1$, the removal of v_1 $(= u_1)$ in $\tilde{A}$ eliminates the two crossings of edges. Therefore the resulting graph is planar and hence has no K_5-minor, by Lemma 11. Thus, $\tilde{A}$ has no K_6-minor when $m = 2$ (since $\tilde{A}$ is an apex graph).

Suppose $m \geq 3$. Cutting A along D_2, we get two annulus triangulations (A', D_1, D_2) and (A'', D_2, D_m) with 2- and $(m-1)$-equator 4-cycle systems, respectively. Let $\tilde{A}'$ (resp., $\tilde{A}''$) be the graph obtained from A' (resp., A'') by adding four edges $v_1v_3, v_2v_4, w_1w_3, w_2w_4$ (resp., $u_1u_3, u_2u_4, w_1w_3, w_2w_4$), where $C_2 = w_1w_2w_3w_4$ in A. By induction hypothesis, $\tilde{A}'$ and $\tilde{A}''$ have no K_6-minors. Applying a 4-sum, we get $\tilde{A} \cup \{w_1w_3, w_2w_4\}$, which has no K_6-minor, by Lemma 12. Therefore, since $\tilde{A}$ is its subgraph, $\tilde{A}$ has no K_6-minor. □

Proof of necessity. Suppose that a triangulation G on $\mathbb{N}_2$ has no K_6-minor. Applying edge contractions to G, we obtain an irreducible triangulation T. Since G has no K_6-minor, neither does T. Then, by Lemma 9, T is isomorphic to $Kh25$ or has an equator 3-cycle. By Lemmas 10 and 13, T has two equator 4-cycles C_1 and C_2 such that the Möbius triangulation cut off from G by C_i has K_4 induced by $V(C_i)$, for $i = 1, 2$. We call these *two K_4s* for a short notation.

CLAIM 2. *G has two K_4s.*

Proof. Suppose that G does not have two K_4s. Let $G = T_0, T_1, \ldots, T_m = T$ be a sequence of triangulations on $\mathbb{N}_2$ such that T_{i+1} is obtained from T_i by a single edge contraction, for $i = 0, \ldots, m-1$. As mentioned above, since T has two K_4s, there exists k such that T_{k+1} has two K_4s but T_k does not. Since K_4 is 3-regular, we may suppose that in T_k, one edge of the K_4 induced by $V(C_1)$ is subdivided, where C_1 is one of the two equator 4-cycles of T_{k+1}. Let C_2 be the equator 4-cycle of T_{k+1} other than C_1. Now cutting along C_2, replacing the Möbius triangulation with boundary C_2 by a 2-cell D, putting a new vertex in D, and joining it to vertices of C_2, we obtain a triangulation T' on $\mathbb{N}_1$. Since each of the two 4-patches in K_4s with nodes $V(C_1)$ has at least one inner vertex, T' is a minor of T_k and has no K_4-quadrangulation. Hence, by Theorem 5, T' has a K_6-minor. Since T' is a minor of G, G has a K_6-minor, a contradiction. □

By Claim 2, if a triangulation G on $\mathbb{N}_2$ has no K_6-minor, then G has two K_4s induced by $V(C_1)$ and $V(C_2)$, where C_1 and C_2 are two equator 4-cycles of G. Hence

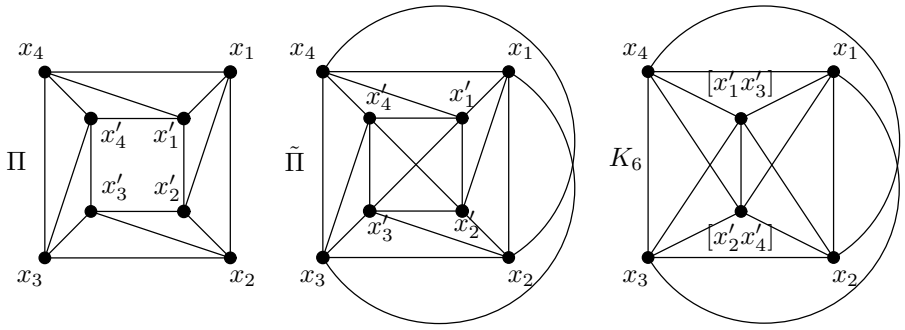


FIG. 10. Π, $\tilde{\Pi}$, *and* K_6.

we shall determine the structure of the annulus triangulation (A, C_1, C_2) obtained from G by removing the Möbius triangulations with boundaries C_1 and C_2.

By Lemma 13, we may suppose that A has no equator 3-cycle. Therefore we can take four disjoint paths P_1, P_2, P_3, P_4 from $V(C_1)$ and $V(C_2)$, by Menger's theorem. We put $C_1 = v_1v_2v_3v_4$, $C_2 = u_1u_2u_3u_4$, and suppose that for $i = 1, 2, 3, 4$, P_i joins v_i and u_i. If A does not have a structure described in (ii)(b) of Theorem 7, then A has two equator 4-cycles D and D' such that the annulus triangulation (A', D, D') has no equator 3- and 4-cycles except D and D', where we put $D = x_1x_2x_3x_4$ and $D' = x'_1x'_2x'_3x'_4$, and suppose that P_i starts at v_i, passes through x_i, x'_i in this order, and ends at u_i, for $i = 1, 2, 3, 4$. Then A' satisfies all of the requirements of Lemma 16, and hence A' has disjoint $x_i - x'_{i+1}$ paths (or $x'_i - x_{i+1}$ paths) for all $i = 1, 2, 3, 4$. Contracting edges on the subpaths of P_i between v_i and x_i, and x'_i and u_i, and edges of A' suitably, we obtain an *antiprism* with eight vertices, denoted by Π, which is obtained from two cycles $x_1x_2x_3x_4$ and $x'_1x'_2x'_3x'_4$ of length 4 and joining x_i to x'_i and x'_{i+1} for each i. (See the left of Figure 10.) Let $\tilde{\Pi}$ be the graph obtained from Π by adding four edges $x_1x_3, x_2x_4, x'_1x'_3, x'_2x'_4$. Then, contracting two edges $x'_1x'_3, x'_2x'_4$, we can transform $\tilde{\Pi}$ into K_6. Since G contains $\tilde{\Pi}$ as a minor, G has a K_6-minor. Hence A must have an equator 4-cycle system. □

5. Remark. Although Hadwiger's conjecture for K_6-minors is proved by Robertson, Seymour, and Thomas [12], the proof is quite lengthy. In this section, we give a quick proof for the case where a given graph triangulates the Klein bottle. One may be able to prove our setting with much careful case analysis. But on the other hand, we do not see yet a very simple way to do it (within a half page proof), so we just put our argument here, using our main result.

PROPOSITION 17. *Let n be a natural number. Every triangulation on the Klein bottle containing no K_n-minor is $(n-1)$-colorable.*

For proving Proposition 17, we use the following lemma.

LEMMA 18 (see [11]). *Let G be a plane graph whose outer cycle C is a 4-cycle and each of whose inner faces is triangular. A 4-coloring of C using precisely four colors extends into a 5-coloring of G.*

Proof of Proposition 17. It is easy to see that every triangulation on a nonspherical surface has a K_n-minor for all n with $1 \le n \le 5$. Moreover, every triangulation on $\mathbb{N}_2$ is known to be 6-colorable [4], and hence the proposition obviously holds when $n \ge 7$. Therefore, let us prove the case for $n = 6$.

By Theorem 7, if a triangulation G on $\mathbb{N}_2$ has no K_6-minor, then G has two equator 4-cycles $C_1 = v_1v_2v_3v_4$ and $C_2 = u_1u_2u_3u_4$ such that each C_i cuts off a

Möbius triangulation (M_i, C_i) with $[V(C_i)] = K_4$. Let A be the remaining annulus triangulation. We first prove that A has a 5-coloring such that each of C_1 and C_2 is colored by exactly four colors.

If C_1 and C_2 have an intersection in A, then we may suppose $u_1 = v_1$. On the other hand, if C_1 and C_2 are disjoint, then we may suppose that v_1 and u_1 are not adjacent. (If such a nonadjacent pair of vertices cannot be chosen, then A must contain a $K_{4,4}$ with partite sets $V(C_1)$ and $V(C_2)$. Since $K_{4,4}$ is not planar, this is impossible, a contradiction.) Add edges v_2v_4 and u_2u_4 to transform A into a triangulation, denoted by A', on the sphere. By four color theorem, A' has a 4-coloring such that v_2, v_3, v_4, and u_2, u_3, u_4 have distinct three colors, respectively. Now, since v_1 and u_1 are identical or nonadjacent, we can recolor v_1 and u_1 by a fifth color, and hence we get a required 5-coloring of A.

Now it suffices to show that M_1 (and M_2) is 5-colorable. We first color v_1, v_2, v_3, v_4 by four colors. Second, extend the 4-coloring to a 5-coloring of the two plane graphs with boundary 4-cycles $v_1v_2v_4v_3$ and $v_2v_3v_1v_4$, using Lemma 18.

In each of M_1, M_2, A, the four vertices of the boundary 4-cycles are colored by distinct four colors, and hence we can get a 5-coloring of G from these 5-colorings of M_1, M_2, A after possibly permuting colors. Therefore, the proposition holds. □

REFERENCES

[1] D. Barnette, *Generating the triangulations of the projective plane*, J. Combin. Theory Ser. B, 33 (1982), pp. 222–230.

[2] P. C. Bonnington and A. Nakamoto, *Geometric realization of a triangulation on the projective plane with one face removed*, Discrete Comp. Geom., 40 (2008), pp. 141–157.

[3] M. DeVos, R. Hedge, K. Kawarabayashi, S. Norine, R. Thomas, and P. Wollan, *K_6-Minors in Large 6-Connected Graphs*, preprint.

[4] G. A. Dirac, *Map-colour theorems*, Canad. J. Math., 4 (1952), pp. 480–490.

[5] G. Fijavź and B. Mohar, *K_6-minors in projective planar graphs*, Combinatorica, 23 (2003), pp. 453–465.

[6] H. Hadwiger, *Über eine Klassifikation der Streckenkomplexe*, Vierteljschr. Naturforsch. Ges. Zürich, 88 (1943), pp. 133–142.

[7] L. K. Jørgensen, *Contractions to K_8*, J. Graph Theory, 18 (1994), pp. 431–448.

[8] S. A. Lawrencenko, *The irreducible triangulations on the torus*, Ukrain. Geom. Sb., 30 (1987), pp. 52–62.

[9] S. Lawrencenko and S. Negami, *Irreducible triangulations of the Klein bottle*, J. Combin. Theory Ser. B, 70 (1997), pp. 265–291.

[10] W. Mader, *Homomorphiesätze für graphen*, Math. Ann., 178 (1968), pp. 154–168.

[11] R. Mukae and A. Nakamoto, *K_6-minors in triangulations and complete quadrangulations*, J. Graph Theory, to appear.

[12] N. Robertson, P. D. Seymour, and R. Thomas, *Hadwiger's conjecture for K_6-free graphs*, Combinatorica, 13 (1993), pp. 279–361.

[13] E. Steinitz and H. Rademacher, *Vorlesungen über die Theorie der Polyeder*, Springer, Berlin, 1934.

[14] T. Sulanke, *Note on the irreducible triangulations of the Klein bottle*, J. Combin. Theory Ser. B, 96 (2006), pp. 964–972.

[15] C. Thomassen, *Planarity and duality of finite and infinite graphs*, J. Combin. Theory Ser. B, 29 (1980), pp. 244–271.

[16] K. Wagner, *Über eine eigenschaft der ebenen komplexe*, Math. Ann., 114 (1937), pp. 570–590.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 109–122

MINIMIZING SONET ADMs IN UNIDIRECTIONAL WDM RINGS WITH GROOMING RATIO SEVEN*

CHARLES J. COLBOURN†, HUNG-LIN FU‡, GENNIAN GE§, ALAN C. H. LING¶, AND HUI-CHUAN LU‖

Abstract. In order to reduce the number of add-drop multiplexers (ADMs) in SONET/WDM networks using wavelength add-drop multiplexing, certain graph decompositions can be used to form a "grooming" that specifies the assignment of traffic to wavelengths. When traffic among nodes is all-to-all and uniform, the drop cost of such a decomposition is the sum, over all graphs in the decomposition, of the number of vertices of nonzero degree in the graph. The number of ADMs required is this drop cost. The existence of such decompositions with minimum cost, when every pair of sites employs no more than $\frac{1}{7}$ of the wavelength capacity, is determined within an additive constant. Indeed when the number n of sites satisfies $n \equiv 1 \pmod 3$ and $n \neq 19$, the determination is exact; when $n \equiv 0 \pmod 3$, $n \not\equiv 18 \pmod{24}$, and n is large enough, the determination is also exact; and when $n \equiv 2 \pmod 3$ and n is large enough, the gap between the cost of the best construction and the cost of the lower bound is independent of n and does not exceed 4.

Key words. traffic grooming, combinatorial designs, block designs, group-divisible designs, optical networks, wavelength-division multiplexing

AMS subject classifications. 68M10, 68R05

DOI. 10.1137/070709141

1. Introduction.

Traffic grooming in optical (SONET) rings arises from amalgamating C low rate signals onto a higher capacity wavelength [15, 25, 26]; C is the *grooming ratio*. Nodes initiate or terminate traffic on a wavelength using an *add-drop multiplexer* (ADM). Finding the minimum number of ADMs, $A(C, n)$, required in an n-node SONET ring with grooming ratio C, is equivalent to the following problem in graphs [4]: Given a number of nodes n and a grooming ratio C, find a partition of the edges of K_n into subgraphs B_ℓ, $\ell = 1, \ldots, s$, with $|E(B_\ell)| \leq C$ such that $\sum_{1 \leq \ell \leq s} |V(B_\ell)|$ is minimum.

Optimal constructions for given grooming ratio C have been obtained using tools of graph and design theory [9]. Results are known for grooming ratio $C = 3$ [1], $C = 4$ [5, 23], $C = 5$ [3], $C = 6$ [2], $C \leq \frac{1}{6}n(n-1)$ [5], and for large values of C [5]. Related problems have been studied for variable traffic requirements [8, 14, 22, 27, 29], for fixed traffic requirements [1, 3, 4, 5, 15, 21, 23, 24, 25, 28, 30], and in the case of

*Received by the editors November 23, 2007; accepted for publication June 20, 2008; published electronically October 31, 2008.
http://www.siam.org/journals/sidma/23-1/70914.html

†Department of Computer Science and Engineering, Arizona State University, P. O. Box 878809, Tempe, AZ 85287-8809 (colbourn@asu.edu).

‡Department of Applied Mathematics, National Chaio Tung University, Hsin Chu, Taiwan, Republic of China (hlfu@math.nctu.edu.tw). This author's work has been partially funded by NSC-94-2115-M009-017.

§Corresponding author. Department of Mathematics, Zhejiang University, Hangzhou 310027, Zhejiang, People's Republic of China (gnge@zju.edu.cn). This author's work has been partially funded by the National Outstanding Youth Science Foundation of China under grant 10825103, the National Natural Science Foundation of China under grant 10771193, the Zhejiang Provincial Natural Science Foundation of China, and the Program for New Century Excellent Talents in University.

¶Department of Computer Science, University of Vermont, Burlington, VT 05405 (aling@cems.uvm.edu).

‖Center of General Education, National United University, Miaoli, Taiwan, Republic of China (hht0936@seed.net.tw). This author's work has been partially funded by NSC-96-2115-M239-002.

bidirectional rings [10, 13]. The explicit correspondence between grooming and graph decomposition is developed in detail in [1, 11].

In this paper we consider grooming with grooming ratio 7. In section 2 we employ linear programming duality to establish a general lower bound on $A(7,n)$. In section 3 we present a complete solution to the existence problem of 4-GDDs of types $24^u m^1$ and $84^u m^1$, which will be used for recursion in subsequent sections. In section 4 we determine $A(7,n)$ with the possible exception of $n=19$ when $n \equiv 1 \pmod 3$. When $n \equiv 0 \pmod 3$ (section 5) we determine $A(7,n)$ with finitely many possible exceptions except when $n \equiv 18 \pmod{24}$; in the latter case we establish a construction whose cost exceeds the lower bound by 1. When $n \equiv 2 \pmod 3$ (section 6) we develop a set of constructions to establish that, with finitely many possible exceptions, the cost does not exceed the lower bound by more than 4, independent of n.

It is natural to ask why the case when $C=7$ is of independent interest. Unlike all cases when $C \le 6$, the graph with the lowest ratio of number of vertices to number of edges does not have C edges; rather it is K_4, a 6-edge graph. This necessitates consideration of decompositions that do not use the minimum number of graphs, and hence determining the minimum number of wavelengths required is quite different than determining the minimum drop cost.

2. The lower bounds. We adapt a strategy using linear programming from [12] that was used in [11] to determine both the cost and the structure of certain optimal groomings. A grooming with ratio 7 is a decomposition of K_n into subgraphs each having at most 7 edges. Its *drop cost*, or just *cost*, is the sum of the numbers of vertices of nonzero degree over all graphs in the decomposition. $A(7,n)$ is the minimum drop cost of a grooming of K_n with grooming ratio 7. Figure 1 displays all of the connected graphs having at most 7 edges. The naming convention is as follows. For each number q of edges and p of vertices, suppose that there are $\gamma_{q,p}$ nonisomorphic graphs. These are named $G_{\ell,q,p}$ for $1 \le \ell \le \gamma_{q,p}$.

In a decomposition, let $\alpha_{\ell,q,p}$ be the number of occurrences of $G_{\ell,q,p}$, and let $\alpha_{q,p} = \sum_{\ell=1}^{\gamma_{q,p}} \alpha_{\ell,q,p}$. Then because every edge appears in exactly one of the chosen subgraphs,

$$\sum_{q=1}^{7}\sum_{p=1}^{8}\sum_{\ell=1}^{\gamma_{q,p}} q \cdot \alpha_{\ell,q,p} = \binom{n}{2}. \tag{1}$$

In order to minimize drop cost, we must compute

$$\min \sum_{q=1}^{7}\sum_{p=1}^{8}\sum_{\ell=1}^{\gamma_{q,p}} p \cdot \alpha_{\ell,q,p}. \tag{2}$$

Figure 1 does not list disconnected graphs, but the cost of a disconnected graph is the sum of the costs of its components, so all feasible decompositions are accounted for. For every graph $G_{\ell,q,p}$, we find that $\frac{p}{q} \ge \frac{2}{3}$. Subtract $\frac{2}{3} \times (1)$ from (2) to restate the minimum drop cost $A(7,n)$ as

$$\frac{n(n-1)}{3} + \min \sum_{q=1}^{7}\sum_{p=1}^{8}\sum_{\ell=1}^{\gamma_{q,p}} \left(p - \frac{2}{3}q\right) \cdot \alpha_{\ell,q,p}. \tag{3}$$

In (3) the triple summation is always nonnegative; it can be zero only when all graphs are isomorphic to K_4. However, structural restrictions can prohibit such a

Graph	$G_{\ell,q,p}$	deg. seq.	$\phi_{\ell,q,p}$	$\psi_{\ell,q,p}$
$G_{1,7,5}$		44222	4	3.5
$G_{2,7,5}$		43322	2.5	2
$G_{3,7,5}$		43331	1	2
$G_{4,7,5}$		33332	1	0.5
$G_{1,7,6}$		442211	4	5
$G_{2,7,6}$		522221	5.5	3.5
$G_{3,7,6}$		422222	5.5	3.5
$G_{4,7,6}$		532211	4	3.5
$G_{5,7,6}$		432221	4	3.5
$G_{6,7,6}$		433211	2.5	3.5
$G_{7,7,6}$		332222	4	2
$G_{1,7,7}$		4421111	4	6.5
$G_{2,7,7}$		5222111	5.5	5
$G_{3,7,7}$		4222211	5.5	5
$G_{4,7,7}$		5321111	4	5
$G_{5,7,7}$		4322111	4	5
$G_{6,7,7}$		4331111	2.5	5
$G_{7,7,7}$		2222222	7	3.5
$G_{8,7,7}$		3222221	5.5	3.5
$G_{9,7,7}$		3322211	4	3.5
$G_{10,7,7}$		3332111	2.5	3.5
$G_{1,7,8}$		71111111	4	8
$G_{2,7,8}$		44111111	4	8
$G_{3,7,8}$		52211111	5.5	6.5
$G_{4,7,8}$		42221111	5.5	6.5
$G_{5,7,8}$		62111111	4	6.5
$G_{6,7,8}$		53111111	4	6.5
$G_{7,7,8}$		43211111	4	6.5
$G_{8,7,8}$		22222211	7	5
$G_{9,7,8}$		32222111	5.5	5
$G_{10,7,8}$		33221111	4	5
$G_{11,7,8}$		33311111	2.5	5
$G_{1,6,4}$		3333	0	0
$G_{1,6,5}$		42222	4.5	3
$G_{2,6,5}$		43221	3	3
$G_{3,6,5}$		33222	3	1.5

Graph	$G_{\ell,q,p}$	deg. seq.	$\phi_{\ell,q,p}$	$\psi_{\ell,q,p}$
$G_{4,6,5}$		33321	1.5	1.5
$G_{1,6,6}$		522111	4.5	4.5
$G_{2,6,6}$		422211	4.5	4.5
$G_{3,6,6}$		432111	3	4.5
$G_{4,6,6}$		222222	6	3
$G_{5,6,6}$		322221	4.5	3
$G_{6,6,6}$		332211	3	3
$G_{7,6,6}$		333111	1.5	3
$G_{1,6,7}$		5211111	4.5	6
$G_{2,6,7}$		4221111	4.5	6
$G_{3,6,7}$		4311111	3	6
$G_{4,6,7}$		6111111	3	6
$G_{5,6,7}$		2222211	6	4.5
$G_{6,6,7}$		3222111	4.5	4.5
$G_{7,6,7}$		3321111	3	4.5
$G_{1,5,4}$		3322	2	1
$G_{1,5,5}$		42211	3.5	4
$G_{2,5,5}$		22222	5	2.5
$G_{3,5,5}$		32221	3.5	2.5
$G_{4,5,5}$		33211	2	2.5
$G_{1,5,6}$		421111	3.5	5.5
$G_{2,5,6}$		511111	3.5	5.5
$G_{3,5,6}$		322111	3.5	4
$G_{4,5,6}$		331111	2	4
$G_{5,5,6}$		222211	5	3
$G_{1,4,4}$		2222	4	2
$G_{2,4,4}$		3221	2.5	2
$G_{1,4,5}$		41111	2.5	5
$G_{2,4,5}$		22211	4	3.5
$G_{3,4,5}$		32111	2.5	3.5
$G_{1,3,3}$		222	3	1.5
$G_{1,3,4}$		2211	3	3
$G_{2,3,4}$		3111	1.5	3
$G_{1,2,3}$		211	2	2.5
$G_{1,1,2}$		11	1	2

FIG. 1. *The graphs.*

selection. In particular, considering the number $\binom{n}{2}$ of edges modulo 6,

$$\sum_{q=1}^{7}\sum_{p=1}^{8}\sum_{\ell=1}^{\gamma_{q,p}}(q \bmod 6)\cdot\alpha_{\ell,q,p} \equiv \begin{cases} 0 \pmod 6 \text{ if } n \equiv 0,1,4,9 \pmod{12}, \\ 1 \pmod 6 \text{ if } n \equiv 2,11 \pmod{12}, \\ 3 \pmod 6 \text{ if } n \equiv 3,6,7,10 \pmod{12}, \\ 4 \pmod 6 \text{ if } n \equiv 5,8 \pmod{12}. \end{cases} \tag{4}$$

We can relax this congruence to linear inequalities. For example, if $n \equiv 3,6,7,10 \pmod{12}$, then

$$\sum_{p=1}^{8}\left[\sum_{\ell=1}^{\gamma_{3,p}}\alpha_{\ell,3,p} + \frac{1}{3}\left(\sum_{q\in\{1,4,7\}}\sum_{\ell=1}^{\gamma_{q,p}}\alpha_{\ell,q,p}\right) + \frac{2}{3}\left(\sum_{q\in\{2,5\}}\sum_{\ell=1}^{\gamma_{q,p}}\alpha_{\ell,q,p}\right)\right] \geq 1, \tag{5}$$

because if there is no graph on three edges, there must be at least three graphs having 1 (mod 3) edges, or one having 1 (mod 3) edges and one having 2 (mod 3) edges.

Every vertex of K_n has degree congruent to $n-1$ mod 3; placing a K_4 in the decomposition does not change this congruence class at any vertex, and hence subgraphs other than K_4 may be needed to accommodate these vertex degrees. Let $\omega_{\ell,q,p}$ be the number of vertices whose degree is congruent to 1 modulo 3 in $G_{\ell,q,p}$, and let $\tau_{\ell,q,p}$ be the number of vertices whose degree is congruent to 2 modulo 3. Now if $n \equiv 0 \pmod 3$, then every vertex has degree 2 modulo 3, and hence at every vertex there must either be a graph itself having degree 2 modulo 3, or two graphs each having degree 1 modulo 3 (there may be more). And if $n \equiv 2 \pmod 3$, then every vertex has degree 1 modulo 3, and hence at every vertex there must either be a graph itself having degree 1 modulo 3, or two graphs each having degree 2 modulo 3. For convenience we write $\phi_{\ell,q,p} = \frac{1}{2}\omega_{\ell,q,p} + \tau_{\ell,q,p}$ and $\psi_{\ell,q,p} = \omega_{\ell,q,p} + \frac{1}{2}\tau_{\ell,q,p}$. These are tabulated for each graph in Figure 1. We conclude that

$$\begin{aligned} &\sum_{q=1}^{7}\sum_{p=1}^{8}\sum_{\ell=1}^{\gamma_{q,p}}\phi_{\ell,q,p}\cdot\alpha_{\ell,q,p} \geq n \quad \text{if} \quad n \equiv 0 \pmod 3, \\ &\sum_{q=1}^{7}\sum_{p=1}^{8}\sum_{\ell=1}^{\gamma_{q,p}}\psi_{\ell,q,p}\cdot\alpha_{\ell,q,p} \geq n \quad \text{if} \quad n \equiv 2 \pmod 3. \end{aligned} \tag{6}$$

THEOREM 2.1. *The cost of an optimal grooming of K_n with grooming ratio 7, $A(7,n)$, is at least*

$$\begin{array}{ll} \frac{2}{3}\binom{n}{2} & \text{if} \quad n \equiv 1,4 \pmod{12}, \\ \frac{2}{3}\binom{n}{2} + 1 & \text{if} \quad n \equiv 7,10 \pmod{12}, \\ \frac{2}{3}\binom{n}{2} + \lceil \frac{n}{12} \rceil & \text{if} \quad n \equiv 0,3,6,15,18,21 \pmod{24}, \\ \frac{2}{3}\binom{n}{2} + \lceil \frac{n}{12} \rceil + 1 & \text{if} \quad n \equiv 9,12 \pmod{24}, \\ \lceil \frac{2}{3}\binom{n}{2} + \frac{2n}{21} \rceil & \text{if} \quad n \equiv 5,8,17 \pmod{21} \\ & \quad \text{or } n \equiv 2,23,32,53,56,77,62,83 \pmod{84}, \\ \lceil \frac{2}{3}\binom{n}{2} + \frac{2n}{21} \rceil + 1 & \text{if} \quad n \equiv 14,35,20,41,44,65,74,11 \pmod{84}. \end{array}$$

Proof. We follow the strategy in [12]. Form a linear program whose variables are

the $\{\alpha_{\ell,q,p}\}$s,

$$\min \sum_{q=1}^{7} \sum_{p=1}^{8} \sum_{\ell=1}^{\gamma_{q,p}} (p - \tfrac{2}{3}q) \cdot \alpha_{\ell,q,p} \tag{7}$$

subject to (4) suitably relaxed, (6), and nonnegativity of each variable.

If $z^\star$ is the minimum, then the cost of any grooming must be at least $\lceil \frac{2}{3}\binom{n}{2} + z^\star \rceil$, since the cost is integral. By forming the dual of (7), any feasible solution to the dual gives a lower bound on all primal feasible solutions, and hence a lower bound on $z^\star$.

Case 1. $n \equiv 1 \pmod 3$: When $n \equiv 1, 4 \pmod{12}$, the linear program is constrained only by nonnegativity, and the dual optimum is 0. When $n \equiv 7, 10 \pmod{12}$, (5) holds. Call its dual variable y_1. An assignment $y_1^\star$ is dual feasible if $y_1^\star \le p - 2$ for every graph $G_{\ell,3,p}$; $y_1^\star \le \frac{3}{2}(p - \frac{2}{3}q)$ for every graph $G_{\ell,q,p}$ with $q \in \{2, 5\}$; and $y_1^\star \le 3(p - \frac{2}{3}q)$ for every graph $G_{\ell,q,p}$ with $q \in \{1, 4, 7\}$. By considering the graphs in Figure 1 the dual optimum of 1 occurs when $y_1^\star = 1$. This raises the lower bound by 1.

Case 2. $n \equiv 0 \pmod 3$: Consider the inequality from (6), and let y_2 be its dual variable. Each graph $G_{\ell,q,p}$ leads to the dual inequality $\phi_{\ell,q,p} y_2 \le p - \frac{2}{3}q$. The dual optimum of $\frac{n}{12}$ arises when $y_2^\star = \frac{1}{12}$; the only graph whose dual inequality is binding is $G_{1,7,5}$ with $\phi_{1,7,5} = 4$ and $5 - \frac{2}{3}7 = \frac{1}{3}$. We can compute the slackness of each variable; for $\alpha_{\ell,q,p}$, the slackness is $p - \frac{2}{3}q - \frac{1}{12}\phi_{\ell,q,p}$. A unit increase in the variable $\alpha_{\ell,q,p}$ increases the dual objective function value by the slackness. The only variables with slackness at most $\frac{1}{2}$ are $\alpha_{2,7,5}$ with slackness $\frac{1}{8}$, $\alpha_{3,7,5}$ and $\alpha_{4,7,5}$ with slackness $\frac{1}{4}$, and $\alpha_{1,5,4}$ with slackness $\frac{1}{2}$. Hence any decomposition of cost less than $\frac{n}{12} + \frac{1}{2}$ consists solely of graphs in $\{G_{\ell,7,5}\}$. To satisfy (6), $\alpha_{7,5} \ge \lceil \frac{n}{4} \rceil$. If $\alpha_{7,5} \ge \frac{n}{4} + \delta$, then adjoining this inequality with dual variable y_3 yields a dual solution $\{y_2 = 0,\ y_3 = \frac{1}{3}\}$ of cost $\frac{n}{12} + \frac{\delta}{3}$, increasing the bound when $\delta \ge 3$. So $\lceil \frac{n}{4} \rceil \le \alpha_{7,5} < \frac{n}{4} + 3$. Because all of the graphs in the decomposition have six or seven edges, $\alpha_{7,5} \equiv 0 \pmod 3$. Thus when $n \equiv 9, 12 \pmod{24}$, $\alpha_{7,5} \equiv 3 \pmod 6$, violating (4). This increases the bound by 1 when $n \equiv 9, 12 \pmod{24}$.

Case 3. $n \equiv 2 \pmod 3$: Again consider the inequality from (6), and let y_2 be its dual variable. Each graph $G_{\ell,q,p}$ leads to the dual inequality $\psi_{\ell,q,p} y_2 \le p - \frac{2}{3}q$. The dual optimum of $\frac{2n}{21}$ arises when $y_2^\star = \frac{2}{21}$; the only graph whose dual inequality is binding is $G_{1,7,5}$ with $\psi_{1,7,5} = \frac{7}{2}$ and $5 - \frac{2}{3}7 = \frac{1}{3}$. We can compute the slackness of each variable; for $\alpha_{\ell,q,p}$, the slackness is $p - \frac{2}{3}q - \frac{2}{21}\psi_{\ell,q,p}$. The only variables with slackness at most $\frac{4}{7}$ are $\alpha_{2,7,5}$ and $\alpha_{3,7,5}$ with slackness $\frac{1}{7}$, $\alpha_{4,7,5}$ with slackness $\frac{2}{7}$, and $\alpha_{1,5,4}$ with slackness $\frac{4}{7}$. An increase of $\frac{4}{7}$ would result in an increase in the integer ceiling when $n \equiv 2, 11, 14, 20 \pmod{21}$, so in these cases we are restricted to K_4s and graphs in $\{G_{\ell,7,5}\}$ to meet the bound. To satisfy (6), $\alpha_{7,5} \ge \lceil \frac{2n}{7} \rceil$. If $\alpha_{7,5} \ge \frac{2n}{7} + \delta$, then adjoining this inequality with dual variable y_3 yields a dual solution $\{y_2 = 0,\ y_3 = \frac{1}{3}\}$ of cost $\frac{2n}{21} + \frac{\delta}{3}$, increasing the bound when $\delta \ge 3$. So $\lceil \frac{2n}{7} \rceil \le \alpha_{7,5} < \frac{2n}{7} + 3$. Because all of the graphs in the decomposition have six or seven edges, $\alpha_{7,5} \equiv 1 \pmod 3$. Thus when $n = 21s + x$ for $x \in \{2, 11, 14, 20\}$, $\alpha_{7,5} = 6s + 1,\ 6s + 4,\ 6s + 4,\ 6s + 7$, respectively. This violates (4) precisely when $n \equiv 44, 65; 11, 74; 14, 35; 20, 41 \pmod{84}$, increasing the bound by 1 in these cases. ∎

We denote by $\mathcal{L}(7, n)$ the lower bound prescribed by Theorem 2.1.

3. Group divisible designs with block size four. A *group divisible design* (GDD) is a triple $(X, \mathcal{G}, \mathcal{B})$, where X is a set of points, $\mathcal{G}$ is a partition of X into *groups*,

and $\mathcal{B}$ is a collection of subsets of X called *blocks* such that any pair of distinct points from X occur together either in some group or in exactly one block, but not both. A K-GDD of type $g_1^{u_1} g_2^{u_2} \cdots g_s^{u_s}$ is a GDD in which every block has size from the set K and in which there are u_i groups of size g_i, $i = 1, 2, \ldots, s$.

A group divisible design $(X, \mathcal{G}, \mathcal{B})$ is *resolvable* if its block set $\mathcal{B}$ admits a partition into *parallel classes*, each parallel class being a partition of the point set X.

A *pairwise balanced design* (PBD) with parameters $(K; v)$ is a K-GDD of type 1^v.

The interested reader may refer to [6, 9] for the undefined terms as well as a general overview of design theory. The main recursive construction that we use is Wilson's fundamental construction (WFC) for GDDs (see, e.g., [9]).

CONSTRUCTION 3.1. *Let $(X, \mathcal{G}, \mathcal{B})$ be a GDD, and let $w : X \to Z^+ \cup \{0\}$ be a weight function on X. Suppose that for each block $B \in \mathcal{B}$, there exists a K-GDD of type $\{w(x) : x \in B\}$. Then there is a K-GDD of type $\{\sum_{x \in G} w(x) : G \in \mathcal{G}\}$.*

A *double group divisible design* (DGDD) is a quadruple $(X, \mathcal{H}, \mathcal{G}, \mathcal{B})$, where X is a set of points, $\mathcal{H}$ and $\mathcal{G}$ are partitions of X (into holes and groups, respectively), and $\mathcal{B}$ is a collection of subsets of X (blocks) such that

(i) for each block $B \in \mathcal{B}$ and each hole $H \in \mathcal{H}$, $|B \cap H| \leq 1$, and
(ii) any pair of distinct points from X which are not in the same hole occur either in some group or in exactly one block, but not both.

A K-DGDD of type $(g_1, h_1^v)^{u_1} (g_2, h_2^v)^{u_2} \cdots (g_s, h_s^v)^{u_s}$ is a DGDD in which every block has size from the set K and in which there are u_i groups of size g_i, each of which intersects each of the v holes in h_i points. (Thus, $g_i = h_i v$ for $i = 1, 2, \ldots, s$. Not every DGDD can be expressed this way, of course, but this is the most general type that we require.) Thus, for example, a *modified group divisible design* (MGDD) K-MGDD of type g^u is a K-DGDD of type $(g, 1^g)^u$. A k-DGDD of type $(g, h^v)^k$ is an incomplete transversal design (ITD) ITD$(k, g; h^v)$ and is equivalent to a set of $k - 2$ holey MOLS of type h^v (see, e.g., [9]). A DGDD is *resolvable* if its block set admits a partition into parallel classes. We use the following existence result.

THEOREM 3.2 (see [20]). *There exists a 4-DGDD of type $(mt, m^t)^n$ if and only if $t, n \geq 4$ and $(t-1)(n-1)m \equiv 0 \pmod 3$ except for $(m, n, t) = (1, 4, 6)$ and except possibly for $m = 3$ and $(n, t) \in \{(6, 14), (6, 15), (6, 18), (6, 23)\}$.*

We also make use of the following simple construction for 4-GDDs.

CONSTRUCTION 3.3 (see [19]). *Suppose that there is a 4-DGDD of type $(g_1, h_1^v)^{u_1} (g_2, h_2^v)^{u_2} \cdots (g_s, h_s^v)^{u_s}$ and that for each $i = 1, 2, \ldots, s$ there is a 4-GDD of type $h_i^v a^1$ where a is a fixed nonnegative integer. Then there is a 4-GDD of type $h^v a^1$, where $h = \sum_{i=1}^s u_i h_i$.*

The following results on transversal designs (TDs) are known.

THEOREM 3.4. *A $TD(k, m)$ exists if*

1. *$k = 5$ and $m \geq 4$ and $m \notin \{6, 10\}$,*
2. *$k = 6$ and $m \geq 5$ and $m \notin \{6, 10, 14, 18, 22\}$,*
3. *$k = 7$ and $m \geq 7$ and $m \notin \{10, 14, 15, 18, 20, 22, 26, 30, 34, 38, 46, 60, 62\}$.*

Finally, we employ the following results on 4-GDDs.

THEOREM 3.5 (see [9, IV.4, Theorem 4.8]). *A 4-GDD of type $3^u m^1$ exists if and only if either $u \equiv 0 \bmod 4$ and $m \equiv 0 \bmod 3$, $0 \leq m \leq (3u - 6)/2$; or $u \equiv 1 \bmod 4$ and $m \equiv 0 \bmod 6$, $0 \leq m \leq (3u - 3)/2$; or $u \equiv 3 \bmod 4$ and $m \equiv 3 \bmod 6$, $0 < m \leq (3u - 3)/2$.*

THEOREM 3.6 (see [17, Theorem 1.7]). *There exists a 4-GDD of type $g^4 m^1$ with $m > 0$ if and only if $g \equiv m \equiv 0 \bmod 3$ and $0 < m \leq \frac{3g}{2}$.*

THEOREM 3.7 (see [18, Theorem 1.6]). *There exists a 4-GDD of type $6^u m^1$ for every $u \geq 4$ and $m \equiv 0 \bmod 3$ with $0 \leq m \leq 3u - 3$ except for $(u, m) = (4, 0)$*

and except possibly for $(u,m) \in \{(7,15),(11,21),(11,24),(11,27),(13,27),(13,33),(17,39),(17,42),(19,45),(19,48),(19,51),(23,60),(23,63)\}$.

THEOREM 3.8 (see [16, Theorem 3.16]). *There exists a 4-GDD of type* $12^u m^1$ *for each* $u \geq 4$ *and* $m \equiv 0 \bmod 3$ *with* $0 \leq m \leq 6(u-1)$.

THEOREM 3.9 (see [16, Theorem 5.21]). *There exists a 4-GDD of type* $2^u m^1$ *for each* $u \geq 6$, $u \equiv 0 \bmod 3$ *and* $m \equiv 2 \bmod 3$ *with* $2 \leq m \leq u-1$ *except for* $(u,m) = (6,5)$ *and except possibly for* $(u,m) \in \{(21,17),(33,23),(33,29),(39,35),(57,44)\}$.

3.1. $g \in \{24, 84\}$.

LEMMA 3.10. *For each* $u \geq 4$, $u \notin \{7,11,13,17,19,23\}$, *there exists a 4-GDD of type* $24^u m^1$ *with* $m \equiv 0 \bmod 3$ *and* $0 \leq m \leq 12(u-1)$.

Proof. For $u = 4$, see Theorem 3.6. For each $u \geq 5$, $u \notin \{7,11,13,17,19,23\}$, take a 4-GDD of type $6^u v^1$ with $v \equiv 0 \bmod 3$ and $0 \leq v \leq 3(u-1)$, and remove the points on the last group of size v; apply weight 4, using 4-MGDDs of type 4^4 and resolvable $\{3\}$-MGDDs of type 4^3, to obtain a $\{3,4\}$-DGDD of type $(24,6^4)^u$ whose triples fall into $3v$ parallel classes. Adjoin $3v$ infinite points to complete the parallel classes, and then fill in 4-GDDs of type $6^u t^1$ with $t \equiv 0 \bmod 3$ and $0 \leq t \leq 3(u-1)$ to obtain a 4-GDD of type $24^u(3v+t)^1$, as desired. □

LEMMA 3.11. *For each* $u \in \{7,11,13,17,19,23\}$, *there exists a 4-GDD of type* $24^u m^1$ *with* $m \equiv 0 \bmod 3$ *and* $3(u-1) \leq m \leq 12(u-1)$.

Proof. For each u, start with a TD$(5,u)$ and adjoin an infinite point ∞ to the groups; then delete a finite point in order to form a $\{5,u+1\}$-GDD of type $4^u u^1$. Note that each block of size $u+1$ intersects the group of size u in the infinite point ∞ and each block of size 5 intersects the group of size u, but certainly not in ∞. Now, in the group of size u, we give ∞ weight 0 or $3(u-1)$ and give the remaining points weight 3, 6, or 9. Give all other points in the $\{5,u+1\}$-GDD weight 6. Replace the blocks in the $\{5,u+1\}$-GDD by 4-GDDs of types 6^u, $6^u(3(u-1))^1$, $6^4 3^1$, $6^4 6^1$, or $6^4 9^1$ to obtain the 4-GDDs, as desired. □

LEMMA 3.12. *For each* $u \in \{7,11,13,17,19,23\}$, *there exists a 4-GDD of type* $24^u m^1$ *with* $m \equiv 0 \bmod 3$ *and* $0 \leq m \leq 3(u-2)$.

Proof. Starting from a 4-DGDD of type $(24,6^4)^u$ coming from Theorem 3.2 and applying Construction 3.3 with 4-GDDs of type $6^u m^1$ to fill in holes, we obtain most of the designs except for $(u,m) \in \{(7,15),(11,21),(11,24),(11,27),(13,27),(13,33),(17,39),(17,42),(19,45),(19,48),(19,51),(23,60),(23,63)\}$.

For the remaining choices for (u,m), take a 4-GDD of type $6^u 3^1$ and remove the points of the last group of size 3; apply weight 4, using 4-MGDDs of type 4^4 and resolvable $\{3\}$-MGDDs of type 4^3, to obtain a $\{3,4\}$-DGDD of type $(24,6^4)^u$ whose triples fall into 9 parallel classes. Adjoin $m-9$ infinite points to complete the parallel classes and then fill in 4-GDDs of type $6^u(m-9)^1$. □

Combining Lemmas 3.10–3.12, we have the following theorem.

THEOREM 3.13. *There exists a 4-GDD of type* $24^u m^1$ *for each* $u \geq 4$ *and* $m \equiv 0 \bmod 3$ *with* $0 \leq m \leq 12(u-1)$.

THEOREM 3.14. *There exists a 4-GDD of type* $84^u m^1$ *for each* $u \geq 4$ *and* $m \equiv 0 \bmod 3$ *with* $0 \leq m \leq 42(u-1)$.

Proof. The proof is similar to that of Lemma 3.10. For each u, take a 4-GDD of type $12^u v^1$ with $v \equiv 0 \bmod 3$ and $0 \leq v \leq 6(u-1)$, and remove the points on the last group of size v; apply weight 7, using 4-MGDDs of type 7^4 and resolvable $\{3\}$-MGDDs of type 7^3, to obtain a $\{3,4\}$-DGDD of type $(84,12^7)^u$ whose triples fall into $6v$ parallel classes. Adjoin $6v$ infinite points to complete the parallel classes, and

then fill in 4-GDDs of type $12^u t^1$ with $t \equiv 0 \bmod 3$ and $0 \le t \le 6(u-1)$ to obtain a 4-GDD of type $84^u(6v+t)^1$, as desired. □

4. Constructions: $n \equiv 1 \pmod 3$. We settle the small cases first.

LEMMA 4.1. *$A(7,n) = \mathcal{L}(7,n)$ for $n \in \{4,7\}$.*

Proof. The lower bound is met for $n = 4$ by a single K_4. The lower bound is realized when $n = 7$: Let $V = \{\infty\} \cup \{0,\ldots,5\}$, and form the three $G_{1,7,5}$s $\{\{i,i+3\}, \{i,i+1\}, \{i,i+4\}, \{i+1,i+3\}, \{i+3,i+4\}, \{\infty,i\}, \{\infty,i+3\}\}$ for $i \in \{0,1,2\}$, arithmetic modulo 6. □

LEMMA 4.2. *$A(7,10) = \mathcal{L}(7,10) + 1 = 32$.*

Proof. The lower bound of 31 is not met. To see this, the only primal variables with slackness at most $\frac{1}{3}$ are for $\{G_{\ell,7,5}\}$. But $6x + 7y = 45$ and $4x + 5y = 31$ admits only the solution $x = 4$ and $y = 3$, i.e., four K_4s and three graphs from $\{G_{\ell,7,5}\}$. There is a unique way to place four K_4s in a K_{10}, and its complement does not partition into three graphs from $\{G_{\ell,7,5}\}$. To produce a decomposition of cost 32, on the 10 points $\{0,\ldots,9\}$ form K_4s on $\{0,1,2,3\}$ and $\{0,4,5,6\}$, and form the graphs

$$\begin{array}{ll}
G_{2,7,5} & \{\{2,4\},\{2,5\},\{2,7\},\{2,9\},\{4,7\},\{5,7\},\{4,9\}\},\\
G_{3,7,5} & \{\{3,9\},\{5,9\},\{6,9\},\{7,9\},\{3,6\},\{3,7\},\{6,7\}\},\\
G_{4,7,5} & \{\{3,4\},\{3,5\},\{3,8\},\{4,8\},\{5,8\},\{1,4\},\{1,5\}\},\\
G_{4,7,5} & \{\{0,7\},\{0,8\},\{0,9\},\{7,8\},\{8,9\},\{1,7\},\{1,9\}\},\\
G_{1,5,4} & \{\{1,8\},\{1,6\},\{2,8\},\{2,6\},\{6,8\}\}.
\end{array}$$ □

LEMMA 4.3. *$\mathcal{L}(7,19) + 1 \le A(7,19) \le \mathcal{L}(7,19) + 2 = 117$.*

Proof. The lower bound of 115 cannot be met. A maximum packing on 19 points has 25 K_4s [7]. Consider the linear program using (5). Using slackness, the only way to achieve a dual objective value of 1 in such a way that at least $21 = \binom{19}{2} - 25 \cdot 6$ edges do not appear in K_4s is to use three graphs in $\{G_{\ell,7,5}\}$. There are 249 nonisomorphic graphs that can be left by a maximum packing of 25 K_4s in K_{19} [2]. $G_{3,7,5}$ cannot be used because it contains a K_4, and the 25 K_4s form a maximum packing. Of the 249 graphs, 79 have degree sequence 3^{14}, 122 have degree sequence $6^1 3^{12}$, and 48 have degree sequence $6^2 3^{10}$. In order to use a $G_{1,7,5}$ there must be at least five vertices of degree 6 or larger; and for $G_{2,7,5}$ there must be at least three. Hence both are ruled out and the only possibility is three $G_{4,7,5}$s. This case can be eliminated by a simple computer search. Thus the drop cost cannot be 115. A solution with drop cost 117 follows:

24 K_4's:	$\{0,1,2,4\}, \{0,3,5,6\}, \{0,7,8,9\}, \{0,10,11,12\}, \{0,13,14,15\},$ $\{0,16,17,18\}, \{1,3,7,10\}, \{1,5,8,11\}, \{1,6,13,16\}, \{1,9,14,17\},$ $\{1,12,15,18\}, \{2,3,8,15\}, \{2,5,9,18\}, \{2,6,10,17\}, \{2,7,12,13\},$ $\{2,11,14,16\}, \{3,4,14,18\}, \{3,9,12,16\}, \{4,5,12,17\}, \{4,6,9,15\},$ $\{5,10,15,16\}, \{6,7,11,18\}, \{6,8,12,14\}, \{8,10,13,18\}$
one $G_{2,7,5}$:	$\{\{3,11\},\{3,13\},\{3,17\},\{11,15\},\{11,17\},\{13,17\},\{15,17\}\}$
two $G_{4,7,5}$:	$\{\{4,7\},\{4,8\},\{4,16\},\{7,16\},\{7,17\},\{8,16\},\{8,17\}\}$ and $\{\{4,10\},\{4,11\},\{4,13\},\{9,10\},\{9,11\},\{9,13\},\{11,13\}\}$
one $G_{7,6,6}$:	$\{\{5,7\},\{5,13\},\{5,14\},\{7,14\},\{7,15\},\{10,14\}\}$. □

THEOREM 4.4. *When $n \equiv 1 \pmod 3$ and $n \notin \{10,19\}$, $A(7,n) = \mathcal{L}(7,n)$. Moreover, $A(7,10) = \mathcal{L}(7,10) + 1$ and $\mathcal{L}(7,19) + 1 \le A(7,19) \le \mathcal{L}(7,19) + 2$.*

Proof. When $n \equiv 1,4 \pmod{12}$, there is a 4-GDD of type 1^n with drop cost $\mathcal{L}(7,n)$. When $n \equiv 7,10 \pmod{12}$ and $n \notin \{10,19\}$, there is a 4-GDD of type $1^{n-7}7^1$ [7]; fill the hole with a solution from Lemma 4.1. □

5. Constructions: $n \equiv 0 \pmod 3$. The lower bound is met for $n = 3$ by a single K_3.

LEMMA 5.1. $A(7,6) = \mathcal{L}(7,6) + 1 = 12$.

Proof. The lower bound of 11 is not met. A decomposition of cost 12 can be produced as follows:

$$\begin{array}{ll} G_{2,7,5} & \{\{0,1\},\{0,2\},\{0,4\},\{0,5\},\{1,4\},\{1,5\},\{2,4\}\}, \\ G_{2,7,5} & \{\{1,2\},\{1,3\},\{2,3\},\{2,5\},\{3,4\},\{3,5\},\{4,5\}\}, \\ G_{1,1,2} & \{\{0,3\}\}. \end{array}$$ □

LEMMA 5.2. $A(7,9) = \mathcal{L}(7,9) + 1 = 27$.

Proof. The lower bound of 26 is not met for $n = 9$ as follows. There can be at most three K_4s on nine points. If there are zero, then at least six graphs are needed, each having a slackness of at least $\frac{1}{3}$; because the total increase in the dual objective function is 2, all of the graphs must be from $\{G_{\ell,7,5}\}$ and cannot account for 36 edges. In the same manner, with one K_4, 30 edges must be accounted for by graphs in $\{G_{\ell,7,5}\}$, each with a slackness of $\frac{1}{3}$ and $G_{1,5,4}$ with a slackness of $\frac{2}{3}$; again this is not possible as 25 is not a multiple of 7. There remain cases with two or three K_4s; each can be eliminated by an exhaustive search.

A decomposition of cost 27 using graphs on at most six edges is given in [2]. We give a different solution here:

$$\begin{array}{ll} G_{1,7,5} & \{\{0,7\},\{0,8\},\{1,7\},\{1,8\},\{2,7\},\{2,8\},\{7,8\}\}, \\ G_{4,7,5} & \{\{0,4\},\{0,5\},\{0,6\},\{1,4\},\{1,5\},\{1,6\},\{4,5\}\}, \\ G_{4,7,5} & \{\{2,4\},\{2,5\},\{2,6\},\{3,4\},\{3,5\},\{3,6\},\{4,6\}\}, \\ G_{4,7,5} & \{\{4,7\},\{4,8\},\{5,6\},\{5,7\},\{5,8\},\{6,7\},\{6,8\}\}, \\ G_{1,6,4} & \{\{0,1\},\{0,2\},\{0,3\},\{1,2\},\{1,3\},\{2,3\}\}, \\ G_{1,2,3} & \{\{3,7\},\{3,8\}\}. \end{array}$$ □

LEMMA 5.3. $A(7,15) = \mathcal{L}(7,15) = 72$.

Proof. Start with a Kirkman triple system of order 9 on $\{0,\ldots,8\}$, in which the first parallel class is $\{B_0, B_1, B_2\}$. Then adjoin points $\{x_0, x_1, x_2, y_0, y_1, y_2\}$. Form nine K_4s by adding y_i to each block of the $(i+2)$nd parallel class. For $i \in \{0,1,2\}$, form a K_4 on $\{x_{i+2}\} \cup B_i$ and a $G_{1,7,5}$ in which the degree 4 vertices are x_i and x_{i+1} and the degree 2 vertices are the elements of B_i. Form a K_4 on $\{x_2, y_0, y_1, y_2\}$. What remains is a $G_{3,6,5}$. □

LEMMA 5.4. $A(7,18) \le \mathcal{L}(7,18) + 1 = 105$.

Proof. Form a 4-GDD of type 3^5 with groups $\{B_j : j = 0,1,2,3,4\}$. Then adjoin points $\{x_0, x_1, x_2\}$. For $i \in \{0,1,2\}$, form a $G_{1,7,5}$ by using the edge $\{x_i, x_{i+1 \bmod 3}\}$, join these vertices to each vertex in B_i, and form a K_4 by adding $x_{i+2 \bmod 3}$ to B_i. For $i \in \{3,4\}$, form a $G_{3,6,5}$ by joining the vertices x_0 and x_1 to vertices in B_i, and form a K_4 by adding x_2 to B_i. This decomposition is of cost 105. □

LEMMA 5.5. $A(7,24) = \mathcal{L}(7,24) = 186$.

Proof. We give the solution on $\{0,1,2,3,4,5,6,7\} \times \mathbb{Z}_3$, writing element (i,j) as i_j.

$$\begin{array}{llll} (0_0,0_1,1_0,4_2), & (0_0,1_1,5_0,6_1), & (0_0,2_0,3_1,3_2), & (0_0,2_1,5_1,5_2), \\ (0_0,2_2,7_0,7_2), & (0_0,6_0,6_2,7_1), & (1_0,1_1,2_1,7_0), & (1_0,2_2,5_1,6_1), \\ (1_0,3_1,5_0,7_1), & (1_0,3_2,4_1,6_2), & (2_0,2_1,4_2,6_1), & (3_0,5_0,6_2,7_2), \\ (4_0,4_1,5_2,7_2), & (3_0,4_0:0_0,1_0,2_0), & (3_0,4_1:5_1,6_1,7_1). & \end{array}$$

The latter two orbits are graphs isomorphic to $G_{1,7,5}$. □

THEOREM 5.6. *$A(7,n) = \mathcal{L}(7,n)$ when $n \equiv 0 \pmod 3$, $n \not\equiv 18 \pmod{24}$, and*

1. $n \geq 96$ *when* $n \equiv 0, 3, 6, 9, 15 \pmod{24}$;
2. $n \geq 276$ *when* $n \equiv 12 \pmod{24}$;
3. $n \geq 309$ *when* $n \equiv 21 \pmod{24}$.

$\mathcal{L}(7, n) \leq A(7, n) \leq \mathcal{L}(7, n) + 1$ *when* $n \equiv 18 \pmod{24}$ *and* $n \geq 114$.

Proof. If $m = n \bmod 24 \in \{0, 3, 6, 9, 15, 18\}$ and $n \geq 96$, form a 4-GDD of type $24^{(n-m)/24}m^1$ from Theorem 3.13; place optimal groomings from Lemma 5.5 on each group of size 24 and an optimal grooming of size m on the exceptional group (from Lemmas 5.1, 5.2, or 5.3 when $m = 6$, 9, 15, respectively). When $m = 18$, use the grooming from Lemma 5.4, missing the lower bound by 1. When $m = 6$, reduce the drop cost by 1 by amalgamating the single edge from this grooming with a K_4 of the 4-GDD to form a $G_{3,7,5}$. When $m = 9$, reduce the drop cost by 1 by amalgamating both edges of the $G_{1,3,2}$ of this grooming with K_4s of the 4-GDD to form $G_{3,7,5}$s.

When $m = n \bmod 24 = 12$, form a 4-GDD of type 20^4 and add four infinite points. On each group, together with the four infinite points, place an optimal grooming from Lemma 5.5 aligning a K_4 on the four infinite points. Suppress the duplicate K_4s so produced. This establishes that $\mathcal{L}(7, 84) = A(7, 84)$. Then filling groups in a 4-GDD of type 24^t84^1 establishes that $A(7, 24t + 84) = \mathcal{L}(7, 24t + 84)$ when $t \geq 8$, i.e., for all $n \geq 276$.

When $m = n \bmod 24 = 21$, form a 4-GDD of type 23^4 and add one infinite point. On each group, together with the infinite point, place an optimal grooming from Lemma 5.5. This establishes that $\mathcal{L}(7, 93) = A(7, 93)$. Then filling groups in a 4-GDD of type 24^t93^1 establishes that $A(7, 24t + 93) = \mathcal{L}(7, 24t + 93)$ when $t \geq 9$, i.e., for all $n \geq 309$. □

6. Constructions: $n \equiv 2 \pmod 3$.

LEMMA 6.1. $A(7, n) = \mathcal{L}(7, n)$ *for* $n \in \{5, 8\}$.

Proof. For K_5, note that $G_{1,7,5} \equiv K_5 \setminus K_3$. Partition K_8 is as follows:

$$\begin{array}{ll} G_{1,7,5} & \{\{0,1\},\{0,2\},\{0,3\},\{0,4\},\{1,2\},\{1,3\},\{1,4\}\}, \\ G_{1,7,5} & \{\{6,7\},\{6,2\},\{6,3\},\{6,4\},\{7,2\},\{7,3\},\{7,4\}\}, \\ G_{3,7,5} & \{\{1,5\},\{2,3\},\{2,4\},\{2,5\},\{3,4\},\{3,5\},\{4,5\}\}, \\ G_{4,7,5} & \{\{1,6\},\{1,7\},\{0,5\},\{0,6\},\{0,7\},\{5,6\},\{5,7\}\}. \end{array}$$ □

LEMMA 6.2. $A(7, 11) = \mathcal{L}(7, 11) = 39$.

Proof. Partition K_{11} on $\{\infty_1, \infty_2\} \cup (\mathbb{Z}_3 \times \mathbb{Z}_3)$ as follows. Include the K_4 $\{\infty_2, 0_2, 1_2, 2_2\}$. Form three $G_{2,7,5}$s as $\{\{i_0, (i+1)_1\}, \{i_0, (i+2)_1\}, \{i_0, (i+1)_2\}, \{i_0, (i+2)_2\}, \{(i+1)_1, (i+2)_1\}, \{(i+1)_1, (i+2)_2\}, \{(i+2)_1, (i+1)_2\}\}$ for $i \in \{0, 1, 2\}$. Then include three $G_{3,7,5}$s as $\{\{\infty_1, i_0\}, \{\infty_1, i_1\}, \{\infty_1, i_2\}, \{i_0, i_1\}, \{i_0, i_2\}, \{i_1, i_2\}, \{\infty_2, i_1\}\}$ for $i \in \{0, 1, 2\}$. Include one last $G_{3,7,5}$: $\{\{\infty_1, \infty_2\}, \{\infty_2, 0_0\}, \{\infty_2, 1_0\}, \{\infty_2, 2_0\}, \{0_0, 1_0\}, \{0_0, 2_0\}, \{1_0, 2_0\}\}$. □

LEMMA 6.3. $A(7, 17) \leq \mathcal{L}(7, 17) + 1 = 94$.

Proof. Start with an $S(2, 4, 16)$ on $\mathbb{Z}_{15} \cup \{\infty\}$ with blocks $\{i, i + 1, i + 3, i + 7\}$ for $i \in \mathbb{Z}_{15}$ and $\{\infty, i, i + 5, i + 10\}$ for $i \in \{0, 1, 2, 3, 4\}$. We adjoin a new point α and modify six of the blocks in the first orbit as follows:

Block	Remove	Add
$\{5, 6, 8, 12\}$	$\{8, 12\}$	$\{\alpha, 5\}, \{\alpha, 8\}$
$\{7, 8, 10, 14\}$	$\{8, 14\}$	$\{\alpha, 7\}, \{\alpha, 10\}$
$\{0, 8, 9, 11\}$	$\{0, 8\}$	$\{\alpha, 0\}, \{\alpha, 9\}$
$\{3, 11, 12, 14\}$	$\{12, 14\}$	$\{\alpha, 3\}, \{\alpha, 12\}$
$\{0, 4, 12, 13\}$	$\{0, 12\}$	$\{\alpha, 4\}, \{\alpha, 13\}$
$\{0, 2, 6, 14\}$	$\{0, 14\}$	$\{\alpha, 2\}, \{\alpha, 14\}$

Now add the K_4 on $\{0,8,12,14\}$. Then delete the K_4 on $\{\infty,1,6,11\}$; on $\{\alpha,\infty,1,6,11\}$, place a K_3 and a $G_{1,7,5}$. The result has 14 K_4s, one K_3, and seven graphs in $\{G_{\ell,7,5}\}$. □

LEMMA 6.4. *When $n \equiv 2 \pmod 6$ and $n \geq 14$, $A(7,n) \leq \frac{2}{3}\binom{n}{2} + \frac{n}{6} = \frac{2}{3}\binom{n}{2} + \frac{2n}{21} + \frac{n}{14}$.*

Proof. Write $h = \frac{n}{2}$. When $h \equiv 1 \pmod 3$ and $h \geq 7$, a 4-GDD of type 2^h exists by Theorem 3.9. It has h groups and $\frac{h(h-1)}{3}$ blocks. For each group, choose a distinct block containing one point of the group (this is an easy exercise using systems of distinct representatives). Then adjoin the pair of each group to its corresponding block to obtain a $G_{3,7,5}$. □

LEMMA 6.5. *When $n \equiv 5 \pmod 6$ and $n \geq 23$, $A(7,n) \leq \frac{2}{3}\binom{n}{2} + \frac{2n}{21} + \frac{n+7}{14}$.*

Proof. Write $h = \frac{n-5}{2}$. When $h \equiv 0 \pmod 3$ and $h \geq 9$, a 4-GDD of type $2^h 5^1$ exists by Theorem 3.9. For each group of size 2, choose a distinct block containing one point of the group and adjoin the pair of each group to its corresponding block to obtain a $G_{3,7,5}$. Then fill the group of size 5 using a solution from Lemma 6.1. □

In order to treat larger cases, we now develop a recursion.

LEMMA 6.6. *There exists a decomposition of K_{21} into nine partial parallel classes of K_3s and six $G_{1,7,5}$s.*

Proof. We present a solution on $\{0,1,\ldots,20\}$ with rows as partial parallel classes:

0 2 13	1 12 15	9 14 17	3 10 20	4 5 19	7 11 16	6 8 18
0 18 20	1 2 16	11 17 19	3 12 13	4 7 8	6 9 10	5 14 15
0 1 11	13 17 18	3 9 16	4 12 14	7 10 19	2 5 6	
0 3 5	1 8 17	4 13 16	7 9 20	6 11 15	2 10 14	
0 8 14	1 5 20	2 3 17	4 10 15	6 13 19	11 12 18	
0 9 15	1 13 14	3 18 19	4 6 20	2 7 12	5 8 16	
0 10 16	1 9 19	12 17 20	3 8 15	2 4 11	5 7 18	
0 12 19	1 10 18	15 16 17	6 7 14	2 8 9	11 13 20	
5 10 17	3 11 14	4 9 18	7 13 15	6 12 16	8 19 20.	

The remaining edges partition into six $G_{1,7,5}$s: $\{\{7i+j, 7i+j+2\}, \{7i+j, 7i+4\}, \{7i+j, 7i+5\}, \{7i+j, 7i+6\}, \{7i+j+2, 7i+4\}, \{7i+j+2, 7i+5\}, \{7i+j+2, 7i+6\}\}$ for $j \in \{0,1\}$ and $i \in \{0,1,2\}$. □

We denote by $X(n)$ the excess over the lower bound, i.e., $X(n) = A(7,n) - \mathcal{L}(7,n)$.

THEOREM 6.7. *Let $(V,\mathcal{G},\mathcal{B})$ be a resolvable group-divisible design of type 7^n, in which the blocks of $\mathcal{B}$ are partitioned into parallel classes $\mathcal{P}_1,\ldots,\mathcal{P}_s$, and for $1 \leq i \leq s$ every block of $\mathcal{P}_i$ has size k_i. Suppose that, for $1 \leq i \leq s$, a 4-GDD of type $3^{k_i}\sigma_i^1$ exists, and that $\sum_{i=1}^s \sigma_i > 0$. Then*

$$A\left(7, 21n+8+\sum_{i=1}^{s}\sigma_i\right) \leq \mathcal{L}\left(7, 21n+8+\sum_{i=1}^{s}\sigma_i\right) + X\left(8+\sum_{i=1}^{s}\sigma_i\right).$$

Proof. Suppose, without loss of generality, that $\sigma_1 > 0$. Give weight three to each point of the GDD $(V,\mathcal{G},\mathcal{B})$. For $2 \leq i \leq s$, adjoin σ_i new infinite points, and place a 4-GDD of type $3^{k_i}\sigma_i^1$ on the inflation of each block of $\mathcal{P}_i$ together with these infinite points. Then proceed similarly for $\mathcal{P}_1$, but adding only $\sigma_1 - 1$ infinite points; in the 4-GDD, delete one point in the group of size σ_1 to form a $\{3,4\}$-GDD of type $3^{k_1}(\sigma_1-1)^1$ in which the blocks of size three form a (frame) parallel class on the $3k_i$ points. On each inflation of a group form a copy of the 21-point design from Lemma 6.6. The nine partial parallel classes of blocks of size 3 formed can be

completed to nine parallel classes on the $21n$ points using the triples from the $\{3,4\}$-GDDs. Finally, add nine further infinite points and extend each of the nine parallel classes to K_4s using these infinite points. The resulting design has a hole on the $8+\sum_{i=1}^{s}\sigma_i$ infinite points added in total, which can be filled with a solution of cost $A(7,8+\sum_{i=1}^{s}\sigma_i)$. □

COROLLARY 6.8.

1. $X(92) \le X(29)$.
2. *For* $n \in \{11,14,17,20,23,26,29\}$, $X(84+n) \le X(n)$.
3. *For* $n \in \{14,20,26,32,38,44,50\}$, $X(105+n) \le X(n)$.
4. *For* $29 \le n \le 71$ *and* $n \equiv 2 \pmod 3$, $X(147+n) \le X(n)$.

Proof. Apply Theorem 6.7 using an RTD$(k,7)$ with $k = 3,4,5,7$ as a resolvable GDD of type 7^k with $s=7$ and $k_1 = \cdots = k_7 = k$. □

COROLLARY 6.9.

1. *For* $29 \le n \le 80$ *and* $n \equiv 2 \pmod 3$, $X(168+n) \le X(n)$.
2. *For* $32 \le n \le 92$ *and* $n \equiv 2 \pmod 6$, $X(189+n) \le X(n)$.
3. *For* $41 \le n \le 107$ *and* $n \equiv 5 \pmod 6$, $X(231+n) \le X(n)$.
4. *For* $44 \le n \le 134$ *and* $n \equiv 2 \pmod 6$, $X(273+n) \le X(n)$.
5. *For* $53 \le n \le 164$ *and* $n \equiv 2 \pmod 3$, $X(336+n) \le X(n)$.

Proof. Apply Theorem 6.7 using an RTD$(7,n)$ with $n = 8,9,11,13,16$ as a resolvable GDD of type 7^n with $s=n$ and $k_1 = \cdots = k_{n-1} = 7$ and $k_n = n$. □

THEOREM 6.10. *For* $x \ge 4$, $0 \le m \le 42(x-1)$, $m \equiv 0 \pmod 3$, *and* $r \in \{11,14,17,20,23,26,29\}$,

$$A(7,84x+m+r) \le \mathcal{L}(7,84x+m+r) + X(m+r).$$

Equivalently, $X(84x+m+r) \le X(m+r)$.

Proof. Form a 4-GDD of type $84^x m^1$ from Theorem 3.14. Adjoin r infinite points, and place a solution on each group of size 84 together with the r points, leaving a hole on the r points (from Corollary 6.8(2.)). On the $m+r$ points, place a solution with excess $X(m+r)$. □

THEOREM 6.11. *For* $m \equiv 2 \pmod 3$ *and* $2 \le m \le 83$, $\mathcal{L}(7,84x+m) \le A(7,84x+m) \le \mathcal{L}(7,84x+m)+X(84x+m)$, *where* $X(84x+m)$ *is given in Table* 1 *(using the final bold entry for* $X(84x+m)$ *in the row for* m *when the table does not provide a value). In particular,* $A(7,84x+m) \le \mathcal{L}(7,84x+m)+4$ *when* $84x+m > 1094$.

Proof. Apply Lemmas 6.1, 6.2, and 6.3 for $x=0$ and $m \in \{5,8,11,17\}$; then apply Lemmas 6.4 and 6.5 to provide an upper bound on $X(84x+m)$ in general. Now apply Corollaries 6.8 and 6.9 to improve these upper bounds. Finally, apply Theorem 6.10. □

7. Conclusions. Grooming with ratio 7 corresponds to the smallest ratio C for which optimal groomings do not consist primarily of C-edge graphs. Consequently, optimal grooming focuses on packings with K_4s in this case. Despite this, the structures of the edges not appearing in K_4s appear to exhibit patterns that repeat modulo 12, 24, and 84 when $n \equiv 1,0,2 \pmod 3$, respectively. In the latter case, techniques for constructing optimal groomings in all cases would necessitate the direct construction of many "small" groomings. Therefore in this paper, we have instead found near-optimal groomings in which the construction deviates from the lower bound by a fixed constant independent of n. When $n \equiv 0,1 \pmod 3$, much more complete characterizations are given. Our conjecture is that, with few small exceptions, the lower bound proved here provides the correct cost of an optimal grooming.

TABLE 1
Least excesses for $84x + m$.

	x													
m	0	1	2	3	4	5	6	7	8	9	10	11	12	13
2	1	6	12	18	2	6	6	6	6	6	6	6	6	**2**
5	0	6	12	4	4	6	6	6	6	6	**4**			
8	0	2	2	18	24	**2**								
11	0	0	12	4	**0**									
14	0	0	2	18	**0**									
17	1	1	13	5	**1**									
20	0	0	2	2	**0**									
23	2	2	14	6	**2**									
26	1	1	3	3	**1**									
29	**2**													
32	2	8	2	4	**2**									
35	2	0	2	20	**0**									
38	2	8	2	4	**2**									
41	2	0	2	20	**0**									
44	2	8	2	4	**2**									
47	3	1	3	21	**1**									
50	3	9	3	5	**3**									
53	4	2	2	22	4	**2**								
56	4	10	4	6	**4**									
59	4	2	2	22	4	**2**								
62	4	10	4	6	**4**									
65	4	2	2	2	4	**2**								
68	4	10	4	6	**4**									
71	5	3	3	3	5	**3**								
74	4	10	4	0	4	4	4	4	4	4	4	4	**0**	
77	6	12	4	4	6	6	6	6	6	**4**				
80	5	11	5	1	5	5	5	5	5	5	5	5	**1**	
83	6	12	4	4	6	6	6	6	6	**4**				

REFERENCES

[1] J.-C. BERMOND AND S. CEROI, *Minimizing SONET ADMs in unidirectional WDM ring with grooming ratio* 3, Networks, 41 (2003), pp. 83–86.

[2] J.-C. BERMOND, C. J. COLBOURN, D. COUDERT, G. GE, A. C. H. LING, AND X. MUÑOZ, *Traffic grooming in unidirectional wavelength-division multiplexed rings with grooming ratio* $C = 6$, SIAM J. Discrete Math., 19 (2005), pp. 523–542.

[3] J.-C. BERMOND, C. J. COLBOURN, A. C. H. LING, AND M.-L. YU, *Grooming in unidirectional rings:* $K_4 - e$ *designs*, Discrete Math., 284 (2004), pp. 57–62.

[4] J.-C. BERMOND AND D. COUDERT, *Traffic grooming in unidirectional WDM ring networks using design theory*, in Proceedings of the IEEE ICC, Anchorage, AK, 2003, pp. 1402–1406.

[5] J.-C. BERMOND, D. COUDERT, AND X. MUÑOZ, *Traffic grooming in unidirectional WDM ring networks: The all-to-all unitary case*, in Proceedings of the 7th IFIP Working Conference on Optical Network Design & Modelling – ONDM'03, 2003, pp. 1135–1153.

[6] T. BETH, D. JUNGNICKEL, AND H. LENZ, *Design Theory*, Bibliographisches Institut, Mannheim, Zurich, 1985.

[7] A. E. BROUWER, *Optimal packings of* K_4*'s into a* K_n, J. Combin. Theory Ser. A, 26 (1979), pp. 278–297.

[8] A. L. CHIU AND E. H. MODIANO, *Traffic grooming algorithms for reducing electronic multiplexing costs in WDM ring networks*, IEEE/OSA Journal of Lightwave Technology, 18 (2000), pp. 2–12.

[9] C. J. COLBOURN AND J. H. DINITZ, EDS., *Handbook of Combinatorial Designs*, 2nd ed., CRC/Chapman and Hall, London, 2007.

[10] C. J. COLBOURN AND A. C. H. LING, *Wavelength add-drop multiplexing and minimizing SONET ADMs*, Discrete Math., 261 (2003), pp. 141–156.

[11] C. J. COLBOURN, G. QUATTROCCHI, AND V. R. SYROTIUK, *Grooming for two-period optical networks*, Networks, to appear.

[12] C. J. Colbourn, G. Quattrocchi, and V. R. Syrotiuk, *Lower bounds for graph decompositions via linear programming duality*, Networks, to appear.
[13] C. J. Colbourn and P.-J. Wan, *Minimizing drop cost for SONET/WDM networks with $\frac{1}{8}$ wavelength requirements*, Networks, 37 (2001), pp. 107–116.
[14] R. Dutta and N. Rouskas, *On optimal traffic grooming in WDM rings*, IEEE J. Sel. Areas Commun., 20 (2002), pp. 110–121.
[15] R. Dutta and N. Rouskas, *Traffic grooming in WDM networks: Past and future*, IEEE Network, 16 (2002), pp. 46–56.
[16] G. Ge and A. C. H. Ling, *Group divisible designs with block size four and group type $g^u m^1$ for small g*, Discrete Math., 285 (2004), pp. 97–120.
[17] G. Ge and R. S. Rees, *On group-divisible designs with block size four and group-type $g^u m^1$*, Des. Codes Cryptogr., 27 (2002), pp. 5–24.
[18] G. Ge and R. S. Rees, *On group-divisible designs with block size four and group-type $6^u m^1$*, Discrete Math., 279 (2004), pp. 247–265.
[19] G. Ge, R. S. Rees, and L. Zhu, *Group-divisible designs with block size four and group-type $g^u m^1$ with m as large or as small as possible*, J. Combin. Theory Ser. A, 98 (2002), pp. 357–376.
[20] G. Ge and R. Wei, *HGDDs with block size four*, Discrete Math., 279 (2004), pp. 267–276.
[21] O. Gerstel, R. Ramaswani, and G. Sasaki, *Cost-effective traffic grooming in WDM rings*, IEEE/ACM Trans. Net., 8 (2000), pp. 618–630.
[22] O. Goldschmidt, D. Hochbaum, A. Levin, and E. Olinick, *The SONET edge-partition problem*, Networks, 41 (2003), pp. 13–23.
[23] J. Q. Hu, *Optimal traffic grooming for wavelength-division-multiplexing rings with all-to-all uniform traffic*, J. Opt. Netw., 1 (2002), pp. 32–42.
[24] J. Q. Hu, *Traffic grooming in WDM ring networks: A linear programming solution*, J. Opt. Netw., 1 (2002), pp. 397–408.
[25] E. Modiano and P. Lin, *Traffic grooming in WDM networks*, IEEE Commun. Mag., 39 (2001), pp. 124–129.
[26] A. Somani, *Survivable traffic grooming in WDM networks*, in Broad Band Optical Fiber Communications Technology—BBOFCT, D. K. Gautam, ed., Jalgaon, India, 2001, pp. 17–45.
[27] P-J. Wan, G. Calinescu, L. Liu, and O. Frieder, *Grooming of arbitrary traffic in SONET/WDM BLSRs*, IEEE J. Sel. Areas Commun., 18 (2000), pp. 1995–2003.
[28] J. Wang, W. Cho, V. Vemuri, and B. Mukherjee, *Improved approaches for cost-effective traffic grooming in WDM ring networks: ILP formulations and single-hop and multihop connections*, IEEE/OSA J. Lightwave Tech., 19 (2001), pp. 1645–1653.
[29] X. Yuan and A. Fulay, *Wavelength assignment to minimize the number of SONET ADMs in WDM rings*, in Proceedings of the IEEE ICC, New York, 2002.
[30] X. Zhang and C. Qiao, *An effective and comprehensive approach for traffic grooming and wavelength assignment in SONET/WDM rings*, IEEE/ACM Trans. Net., 8 (2000), pp. 608–617.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 123–135

THE ERDÖS–FALCONER DISTANCE PROBLEM, EXPONENTIAL SUMS, AND FOURIER ANALYTIC APPROACH TO INCIDENCE THEOREMS IN VECTOR SPACES OVER FINITE FIELDS*

ALEX IOSEVICH† AND DOOWON KOH†

Abstract. We study the Erdös–Falconer distance problem in vector spaces over finite fields with respect to the cubic metric. Estimates for discrete Airy sums and Adolphson–Sperber estimates for exponential sums in terms of Newton polyhedra play a crucial role. Similar techniques are used to study the incidence problem between points and cubic and quadratic curves. As a result we obtain a nontrivial range of exponents that appear to be difficult to attain using combinatorial methods.

Key words. distance sets, Newton diagrams, Gauss sums, multiplicative characters, exponential sums, Kloosterman sums

AMS subject classification. 52C10

DOI. 10.1137/060669875

1. Introduction.

1.1. The Erdös distance problem. The Erdös distance conjecture in the Euclidean space says that if E is a finite subset of $\mathbb{R}^d$, $d \geq 2$, then

$$\#\Delta(E) \gtrapprox (\#E)^{\frac{2}{d}}, \tag{1.1}$$

where

$$\Delta(E) = \{|x-y| : x, y \in E\},$$

with $|x-y|^2 = (x_1-y_1)^2 + \cdots + (x_d-y_d)^2$, and here, and throughout this paper, $X \lesssim Y$ means that there exists $C > 0$ such that $X \leq CY$ and $X \lessapprox Y$, with the controlling parameter N, means that for every $\epsilon > 0$ there exists $C_\epsilon > 0$ such that $X \leq C_\epsilon N^\epsilon Y$.

Taking $E = \mathbb{Z}^d \cap [0, N^{\frac{1}{d}}]^d$ shows that (1.1) cannot, in general, be improved. The conjecture has not been solved in any dimension. See, for example, [14], [2], and the references contained therein for the description of the conjecture, background material, and a survey of recent results.

In this paper we study the Erdös distance problem in vector spaces over finite fields. This problem was recently addressed by Tao [19], who relates it to some interesting questions in combinatorics, and, more recently, by Iosevich and Rudnev [9]. We shall describe these results later.

Let $\mathbb{F}_q$ denote the finite field with q elements, and let $\mathbb{F}_q^d$ denote the d-dimensional vector space over this field. Let $E \subset \mathbb{F}_q^d$, $d \geq 2$. Then a possible analogue of the classical Erdös distance problem is to determine the smallest possible cardinality of the set

$$\Delta_n(E) = \{||x-y||_n = (x_1-y_1)^n + \cdots + (x_d-y_d)^n : x, y \in E\},$$

*Received by the editors September 14, 2006; accepted for publication (in revised form) June 28, 2008; published electronically October 31, 2008. This work was supported by the NSF grant DMS04-56306.

http://www.siam.org/journals/sidma/23-1/66987.html

†Mathematics Department, 202 Mathematical Sciences Building, University of Missouri, Columbia, MO 65211 (iosevich@math.missouri.edu, koh@math.missouri.edu).

with n a positive integer ≥ 2, viewed as a subset of $\mathbb{F}_q$.

In the finite field setting, the estimate (1.1) cannot hold without further restrictions. To see this, let $E = \mathbb{F}_q^d$. Then $\#E = q^d$ and $\#\Delta(E) = q$. Furthermore, an interesting feature of the Erdös distance problem in the finite field setting with $n = 2$ is the existence of nontrivial spheres of zero radius. These are sets of the form $\{x \in \mathbb{F}_q^d : x_1^2 + x_2^2 + \cdots + x_d^2 = 0\}$, and several assumptions in the statements of results below are there precisely to deal with issues created by the presence of this object. For example, suppose -1 is a square in $\mathbb{F}_q$. Using spheres of zero radius one can show, in even dimensions, that there exists a set of cardinality precisely $q^{\frac{d}{2}}$ such that all of the distances, $(x_1 - y_1)^2 + \cdots + (x_d - y_d)^2$, are zero. What's more, suppose $\mathbb{F}_q$ is a finite field such that $q = p^2$, where p is a prime. Then $E = \mathbb{F}_p^d$ is naturally embedded in $\mathbb{F}_q^d$, has cardinality $q^{\frac{d}{2}}$, and determines only $\sqrt{q}$ distances. If $n > 2$, then the situation is equally fascinating. For example, if $n = 3$ and $d = 2$, then the equation $x_1^3 + x_2^3 = 0$ always has at least q solutions, since the cube root of -1 is -1. This equation may have as many as $3q$ solutions if the primitive cube root of -1 is in the field.

With these examples as our guide, we generalize the conjecture originally stated in [9] in the case $n = 2$ as follows.

CONJECTURE 1.1. *Let $E \subset \mathbb{F}_q^d$ of cardinality $\geq Cq^{\frac{d}{2}}$, with C sufficiently large. Then*

$$\#\Delta_n(E) \gtrapprox q.$$

The authors of [9] conjecture that the constant C that appears above may be taken to be any number bigger than one, at least in the case $n = 2$. It is interesting to note that if $n > 2$, then the situation becomes more complicated. For example, as we pointed out above, if $n = 3$ and $d = 2$, then the number of points on the curve $x_1^3 + x_2^3 = 0$ may be as high as $3q$, depending on whether or not the primitive cube root of -1 is in the field. Thus a corresponding conjecture in the case $n > 2$ must be designed with these issues in mind.

2. Previous results. A Euclidean plane argument due to Erdös [6] can be applied to the finite field set-up under the assumption of Conjecture 1.1 to show that if $d = 2$ and $\#E \geq Cq$, with C sufficiently large, then

$$\#\Delta_n(E) \gtrsim (\#E)^{\frac{1}{2}}. \tag{2.1}$$

This result was improved by Bourgain, Katz, and Tao [3], who showed using intricate incidence geometry that for every $\epsilon > 0$, there exists $\delta > 0$ such that if $\#E \lesssim q^{2-\epsilon}$, then

$$\#\Delta_2(E) \gtrsim q^{\frac{1}{2}+\delta}.$$

The relationship between ϵ and δ in the above argument is difficult to determine. Moreover, matters are even more subtle in higher dimensions in the context of vector spaces over finite fields, because the intersection of the analogues of spheres, both quadratic and cubic, in $\mathbb{F}_q^d$ may be quite complicated, and the standard induction on the dimension argument in $\mathbb{R}^d$ (see, e.g., [2]) that allows one to bootstrap the estimate (2.1) into the estimate

$$\#\Delta_{\mathbb{R}^d}(E) \gtrsim (\#E)^{\frac{1}{d}} \tag{2.2}$$

does not immediately go through. We establish the finite field analogue of the estimate (2.2) below using Fourier analytic methods and number theoretic properties of Kloosterman sums and its more general analogues.

Another way of thinking of Conjecture 1.1 is in terms of the Falconer distance conjecture [7] in the Euclidean setting which says that if the Hausdorff dimension of a set in $\mathbb{R}^d$ exceeds $\frac{d}{2}$, then the Lebesgue measure of the distance set is positive. Conjecture 1.1 implies that if the size of the set is greater than $q^{\frac{d}{2}}$, then the distance set contains a positive proportion of all the possible distances, an analogous statement.

In [9], the authors proved the following result.

THEOREM 2.1. *Let $E \subset \mathbb{F}_q^d$, $d \geq 2$, such that $\#E \geq Cq^{\frac{d+1}{2}}$. Then if C is sufficiently large, then $\Delta_2(E)$ contains every element of $\mathbb{F}_q$.*

3. Main results of this paper.

3.1. Distances determined by a single set. Our first result is the version of Theorem 2.1 for cubic metrics.

THEOREM 3.1. *Suppose that q is a prime number congruent to 1 modulo 3. Let $E \subset \mathbb{F}_q^d$ such that $\#E \geq Cq^{\frac{d+1}{2}}$. Then if C is sufficiently large, then $\Delta_3(E)$ contains every element of $\mathbb{F}_q$.*

Suppose that $d = 2$, and $n \geq 2$. Then if $\#E \geq Cq^{\frac{3}{2}}$ for C sufficiently large, then $\Delta_n(E)$ contains every element of $\mathbb{F}_q$.

COROLLARY 3.2. *Suppose that q is a prime number congruent to 1 modulo 3. Let $E \subset \mathbb{F}_q^d$, $d \geq 2$, such that $\#E = Cq^{\frac{d+1}{2}}$. Then if C is sufficiently large, then*

$$\#\Delta_3(E) \approx (\#E)^{\frac{2}{d+1}}.$$

In two dimensions, the same conclusion, with $d = 2$, holds for any $n \geq 2$.

Note that in the case $d = 2$, the exponent $\frac{2}{3}$ obtained via the corollary, for the given range of parameters, is a much better exponent than the one obtained by the incidence argument due to Erdös described in (2.1). Also, we point out once more that Erdös' argument does not generalize to higher dimensions, at least not very easily, due to the possibly complicated intersection properties of cubic varieties.

3.2. Szemerédi–Trotter-type incidence theorems and distances between pairs of sets. As in the case $n = 2$, the proof of Theorem 3.1 can be modified to yield a good upper bound on the number of incidences between points and cubic surfaces in vector spaces over finite fields. It is an analogue, and a higher dimensional generalization, of the following classical result due to Szemerédi and Trotter.

THEOREM 3.3. *The number of incidences between N points and M lines (or circles of the same radius) in the plane is*

$$\lesssim N + M + (NM)^{\frac{2}{3}}.$$

Our incident estimate is the following theorem.

THEOREM 3.4. *Suppose that q is a prime number congruent to 1 modulo 3. Let $E, F \subset \mathbb{F}_q^d$, $d \geq 2$. Then if $j \neq 0$, then*

$$\#\{(x,y) \in E \times F : (x_1 - y_1)^3 + \cdots + (x_d - y_d)^3 = j\}$$

$$\lesssim \#E \cdot \#F \cdot q^{-1} + q^{\frac{d-1}{2}} \cdot (\#E)^{\frac{1}{2}} \cdot (\#F)^{\frac{1}{2}}.$$

Similarly, if q is a prime number and $j \neq 0$, then

$$\#\{(x,y) \in E \times F : (x_1 - y_1)^2 + \cdots + (x_d - y_d)^2 = j\}$$
$$\lesssim \#E \cdot \#F \cdot q^{-1} + q^{\frac{d-1}{2}} \cdot (\#E)^{\frac{1}{2}} \cdot (\#F)^{\frac{1}{2}}.$$

In two dimensions, the same result holds, with $d = 2$, with Δ_3 replaced by Δ_n for any $n \geq 2$.

Remark 3.5. In particular, if $\#E \approx \#F \approx q^{\frac{d+1}{2}}$, then the number of incidences between points in E and "spheres," quadratic or cubic, centered at the elements of F is $\lesssim q^d$.

To make the numerology more transparent, Theorem 3.4 says that if $N \approx q^{\frac{d+1}{2}}$, then the number of incidences between $\approx N$ points and $\approx N$ spheres, cubic or quadratic, in $\mathbb{F}_q^d$ is $\lesssim q^d = N^{\frac{2d}{d+1}}$. In two dimensions this says that the number of incidences between N points and N circles is $\lesssim N^{\frac{4}{3}}$, provided that $N \approx q^{\frac{d+1}{2}}$, matching in this setting the exponent in the celebrated result due to Szemerédi and Trotter in the Euclidean plane (see Theorem 3.3).

An easy modification of the method used to prove Theorem 3.4 above yields the following distance set result.

COROLLARY 3.6. *Let $E, F \subset \mathbb{F}_q^d$, $d \geq 2$. Suppose that q is a prime number congruent to 1 modulo 3 and $\#E \cdot \#F \geq Cq^{d+1}$. Let $\Delta_3(E,F) = \{||x-y||_3 : x \in E, y \in F\}$. Then if C is sufficiently large, then $\Delta_3(E,F)$ contains every element of $\mathbb{F}_q^*$.*

As before, in two dimensions the same conclusion holds with $d = 2$ and Δ_3 replaced by $\Delta_n(E)$.

Observe that if $E = F$, then we can safely say that in fact $\Delta_3(E,F)$ contains every element of $\mathbb{F}_q$, but if $E \neq F$, then the zero distance may not be present.

We also call the reader's attention to the fact that an analogous version of this result was independently obtained by Shparlinski in [17].

4. Fourier analytic preliminaries and notation. Let $\mathbb{F}_q$ be a finite field with q elements, where q is a prime number. Let

$$\chi(t) = e^{\frac{2\pi i}{q}t}.$$

Given a complex valued function f on $\mathbb{F}_q^d$, define the Fourier transform of f by

$$\widehat{f}(m) = q^{-d} \sum_{x \in \mathbb{F}_q^d} \chi(-x \cdot m) f(x).$$

We also need the following basic identity, typically known as the Plancherel theorem. Let f be as above. Then

$$\sum_{m \in \mathbb{F}_q^d} |\widehat{f}(m)|^2 = q^{-d} \sum_{x \in \mathbb{F}_q^d} |f(x)|^2.$$

5. Proof of the first part of Theorem 3.1. Let $\chi(s) = e^{\frac{2\pi i}{q}s}$. Let S_j denote the characteristic function of the cubic sphere

$$\{x \in \mathbb{F}_q^d : ||x||_3 = j\},$$

where, as before,

$$||x||_3 = x_1^3 + \cdots + x_d^3.$$

The key estimate of the paper is the following theorem.

THEOREM 5.1. *Let* $||x||_3 = x_1^3 + \cdots + x_d^3$. *Suppose that* q *is a prime number congruent to* 1 *modulo* 3 *and* $j \neq 0$. *Then if* $m \neq (0, \ldots, 0)$, *then*

$$\left|\widehat{S}_j(m)\right| = \left| q^{-d} \sum_{\{x \in \mathbb{F}_q^d : ||x||_3 = j\}} \chi(x \cdot m) \right| \lesssim q^{-\frac{d+1}{2}},$$

and if $m = (0, \ldots, 0)$, *then*

$$\widehat{S}_j(m) = q^{-1} + O(q^{-\frac{d+1}{2}}) \approx q^{-1}.$$

For $j \neq 0$, consider

$$\begin{aligned} &\#\{(x,y) \in E \times E : ||x-y||_3 = j\} \\ &\quad = \sum_{x,y \in \mathbb{F}_q^d} E(x)E(y)S_j(x-y) \\ &= q^{2d} \sum_m \left|\widehat{E}(m)\right|^2 \widehat{S}_j(m) = \mathrm{A} + \mathrm{B}, \end{aligned}$$

where

$$\mathrm{A} = q^{2d} \left|\widehat{E}(0, \ldots, 0)\right|^2 \widehat{S}_j(0, \ldots, 0),$$

and

$$\mathrm{B} = q^{2d} \sum_{m \neq (0,\ldots,0)} \left|\widehat{E}(m)\right|^2 \widehat{S}_j(m).$$

Using the second part of Theorem 5.1,

$$\mathrm{A} \approx q^{2d} q^{-2d} (\#E)^2 \cdot q^{-1}.$$

Whereas using the first part of Theorem 5.1,

$$\begin{aligned} |\mathrm{B}| &\lesssim q^{2d} q^{-\frac{d+1}{2}} \sum_{m \neq (0,\ldots,0)} \left|\widehat{E}(m)\right|^2 \\ &\lesssim q^{2d} q^{-\frac{d+1}{2}} q^{-d} \sum_{x \in \mathbb{F}_q^d} E^2(x) = q^{\frac{d-1}{2}} \cdot \#E. \end{aligned}$$

We therefore obtain that

$$\#\{(x,y) \in E \times E : ||x-y||_3 = j\} = \mathrm{A} + \mathrm{B},$$

where

$$\mathrm{A} \gtrsim (\#E)^2 q^{-1},$$

and

$$|\mathrm{B}| \lesssim \#E \cdot q^{\frac{d-1}{2}}.$$

We conclude that if $\#E \geq Cq^{\frac{d+1}{2}}$, with C sufficiently large, then

$$\#\{(x,y) \in E \times E : ||x-y||_3 = j\} > 0$$

for each $j \neq 0$. This completes the proof of Theorem 3.1. □

6. Proof of Theorem 5.1. We have

$$\widehat{S}_j(m) = q^{-d} \sum_{\{x \in \mathbb{F}_q^d : ||x||_3 = j\}} \chi(-x \cdot m)$$

$$= q^{-1}\delta(m) + q^{-d-1} \sum_x \sum_{t \in \mathbb{F}_q^*} \chi(t(||x||_3 - j))\chi(-x \cdot m),$$

where $\delta(m) = 1$ if $m = (0, \ldots, 0)$ and 0 otherwise.

LEMMA 6.1. *Let χ be a nontrivial additive character of $\mathbb{F}_q$ with q congruent to 1 modulo 3. Suppose that $m = (m_1, \ldots, m_l) \in (\mathbb{F}_q^*)^l$. Then for any multiplicative character ψ of $\mathbb{F}_q$ of order 3 and $t \neq 0$, we have*

$$\prod_{j=1}^{l} \sum_{s_j \in \mathbb{F}_q} \chi(-s_j m_j + s_j^3 t)$$

$$= \psi^{-l}(t) \sum_{s_1, \ldots, s_l \in \mathbb{F}_q^*} \chi(s_1 + \cdots + s_l + m_1^3 t^{-1} s_1^{-1} + \cdots + m_l^3 t^{-1} s_l^{-1}) \psi(s_1) \cdots \psi(s_l),$$

where $3^{-3} m_j^3$ is denoted by m_j^3 in the right-hand side of the equation.

We shall also need the following result due to Duke and Iwaniec [5].

THEOREM 6.2. *Suppose that q is congruent to 1 modulo 3, and let ψ be a multiplicative character of order three. Then*

$$\sum_{s \in \mathbb{F}_q} \chi(as^3 + s) = \sum_{s \in \mathbb{F}_q^*} \psi(sa^{-1}) \chi(s - (3^3 as)^{-1})$$

for any $a \in \mathbb{F}_q^*$.

It follows that

$$\sum_{s \in \mathbb{F}_q} \chi(-sm_j + s^3 t) = \sum_{s \in \mathbb{F}_q} \chi(s - s^3 t m_j^{-3})$$

$$= \sum_{s \in \mathbb{F}_q^*} \psi(st^{-1}) \chi(s + m_j^3 t^{-1} 3^{-3} s^{-1})$$

since ψ is a multiplicative character of $\mathbb{F}_q$ of order three and $m_j \neq 0$. Absorbing 3^{-3} into m_j to make the notations simple, we complete the proof of Lemma 6.1. □

LEMMA 6.3. *Let χ be a nontrivial additive character of $\mathbb{F}_q$ with q congruent to 1 modulo 3. Then for any multiplicative character ψ of $\mathbb{F}_q$ of order 3 and $t \neq 0$, we have*

$$\left(\sum_{s \in \mathbb{F}_q} \chi(ts^3) \right)^l = \sum_{r=0}^{l} \binom{l}{r} q^l \, \psi^{-(l+r)}(t) \Big(\widehat{\psi}(-1) \Big)^{l-r} \Big(\widehat{\psi^2}(-1) \Big)^r,$$

where $\binom{\cdot}{\cdot}$ *is a binomial coefficient , l is a positive integer, and the Fourier transform of a multiplicative character* ψ *of* $\mathbb{F}_q$ *is given by*

$$\widehat{\psi}(v) = q^{-1} \sum_{s \in \mathbb{F}_q^*} \chi(-vs)\psi(s).$$

Remark 6.4. $\widehat{\psi}(v) = O(q^{-\frac{1}{2}})$ for $v \neq 0$.

To prove Lemma 6.3, we need the following theorem. For the proof, see [13, Theorem 5.30, p. 217].

THEOREM 6.5. *Let* χ *be a nontrivial additive character of* $\mathbb{F}_q$, $n \in \mathbb{N}$, *and* ψ *a multiplicative character of* $\mathbb{F}_q$ *of order* $h = gcd(n, q-1)$. *Then*

$$\sum_{s \in \mathbb{F}_q} \chi(ts^n + b) = \chi(b) \sum_{k=1}^{h-1} \psi^{-k}(t)\, G(\psi^k, \chi)$$

for any $t, b \in \mathbb{F}_q$ *with* $t \neq 0$, *where* $G(\psi^k, \chi) = \sum_{s \in \mathbb{F}_q^*} \psi^k(s)\chi(s)$.

By using Theorem 6.5, we see that for any multiplicative character ψ of order three,

$$\left(\sum_{s \in \mathbb{F}_q} \chi(ts^3) \right)^l = \left(\sum_{k=1}^{2} \psi^{-k}(t) \sum_{s \in \mathbb{F}_q^*} \psi^k(s)\, \chi(s) \right)^l$$

$$= \left(\psi^{-1}(t) \sum_{s \in \mathbb{F}_q^*} \psi(s)\, \chi(s) + \psi^{-2}(t) \sum_{s \in \mathbb{F}_q^*} \psi^2(s)\chi(s) \right)^l$$

$$= \Big(G_1(t) + G_2(t) \Big)^l = \sum_{r=0}^{l} \binom{l}{r} G_1(t)^{l-r} G_2(t)^r,$$

where

$$G_1(t) = \psi^{-1}(t) \sum_{s \in \mathbb{F}_q^*} \psi(s)\chi(s),$$

and

$$G_2(t) = \psi^{-2}(t) \sum_{s \in \mathbb{F}_q^*} \psi^2(s)\chi(s).$$

Note that $G_1(t) = q\psi^{-1}(t)\, \widehat{\psi}(-1)$ and $G_2(t) = q\psi^{-2}(t)\, \widehat{\psi^2}(-1)$.

Thus we conclude that

$$\left(\sum_{s \in \mathbb{F}_q} \chi(ts^3) \right)^l = \sum_{r=0}^{l} \binom{l}{r} q^l \psi^{-(l+r)}(t) \Big(\widehat{\psi}(-1) \Big)^{l-r} \Big(\widehat{\psi^2}(-1) \Big)^r.$$

We are now ready to prove Theorem 5.1. First, we assume that $m = (0, \ldots, 0) \in \mathbb{F}_q^d$. Then, using Lemma 6.3, we see that

$$\widehat{S}_j(0, \ldots, 0) = q^{-d} \sum_{\{x \in \mathbb{F}_q^d : ||x||_3 = j\}} 1$$

$$= q^{-1} + q^{-d-1} \sum_{t \in \mathbb{F}_q^*} \chi(-tj) \sum_x \chi(t(||x||_3))$$

$$= q^{-1} + q^{-d-1} \sum_{t \in \mathbb{F}_q^*} \chi(-tj) \sum_{r=0}^{d} \left(\begin{array}{c} d \\ r \end{array} \right) q^d \psi^{-(d+r)}(t) \Big(\widehat{\psi}(-1)\Big)^{d-r} \Big(\widehat{\psi^2}(-1)\Big)^r$$

$$= q^{-1} + q^{-1} \sum_{r=0}^{d} \left(\begin{array}{c} d \\ r \end{array} \right) \Big(\widehat{\psi}(-1)\Big)^{d-r} \Big(\widehat{\psi^2}(-1)\Big)^r \sum_{t \in \mathbb{F}_q^*} \chi(-tj) \psi^{-(d+r)}(t)$$

$$= q^{-1} + q^{-1} \sum_{r=0}^{d} \left(\begin{array}{c} d \\ r \end{array} \right) \Big(\widehat{\psi}(-1)\Big)^{d-r} \Big(\widehat{\psi^2}(-1)\Big)^r \widehat{q\psi^{-(d+r)}}(j)$$

$$= q^{-1} + O(q^{-\frac{d+1}{2}}) \approx q^{-1}.$$

In the last equality, we used the fact that $\widehat{\psi}(v) = O(q^{-\frac{1}{2}})$ for any multiplicative character of $\mathbb{F}_q$ with $v \neq 0$. Thus the second part of Theorem 5.1 is proved.

In order to prove the first part of Theorem 5.1, we shall deal with the problem in case $m = (m_1, \ldots, m_d) \neq (0, \ldots, 0)$. Suppose that $m_j \neq 0$ for $j \in J \subset \{1, 2, \ldots, d\}$ and $m_j = 0$ for $j \in \{1, 2, \ldots, d\} \setminus J = J^{'}$. Without loss of generality, we may assume that $J = \{1, 2, \ldots, l\}$ and $J^{'} = \{l+1, \ldots, d\}$ for some $l = 1, 2, \ldots, d$. Using Lemma 6.1 and Lemma 6.3, we see that

$$\widehat{S_j}(m) = q^{-d-1} \sum_{t \in \mathbb{F}_q^*} \chi(-tj) \sum_{x \in \mathbb{F}_q^d} \chi(t||x||_3 - m \cdot x)$$

$$= q^{-d-1} \sum_{t \in \mathbb{F}_q^*} \chi(-tj) \left(\prod_{k=1}^{l} \sum_{s_k \in \mathbb{F}_q} \chi(ts_k^3 - m_k s_k) \right) \left(\prod_{k=l+1}^{d} \sum_{s_k \in \mathbb{F}_q} \chi(ts_k^3) \right)$$

$$= q^{-d-1} \sum_{t \in \mathbb{F}_q^*} \chi(-tj) \psi^{-l}(t) \sum_{s_1, \ldots, s_l \in \mathbb{F}_q^*}$$

$$\times \chi(s_1 + \cdots + s_l + m_1^3 t^{-1} s_1^{-1} + \cdots + m_l^3 t^{-1} s_l^{-1}) \psi(s_1) \cdots \psi(s_l)$$

$$\times \sum_{r=0}^{d-l} \left(\begin{array}{c} d-l \\ r \end{array} \right) q^{d-l} \psi^{-(d-l+r)}(t) \Big(\widehat{\psi}(-1)\Big)^{d-l-r} \Big(\widehat{\psi^2}(-1)\Big)^r$$

$$= q^{-1-l} \sum_{r=0}^{d-l} \left(\begin{array}{c} d-l \\ r \end{array} \right) \Big(\widehat{\psi}(-1)\Big)^{d-l-r} \Big(\widehat{\psi^2}(-1)\Big)^r \sum_{t \in \mathbb{F}_q^*} \chi(-tj) \psi^{-(d+r)}(t)$$

$$\times \sum_{s_1, \ldots, s_l \in \mathbb{F}_q^*} \chi(s_1 + \cdots + s_l + m_1^3 t^{-1} s_1^{-1} + \cdots + m_l^3 t^{-1} s_l^{-1}) \psi(s_1) \cdots \psi(s_l).$$

Since $\binom{d-l}{r} \Big(\widehat{\psi}(-1)\Big)^{d-l-r} \Big(\widehat{\psi^2}(-1)\Big)^r = O(q^{-\frac{1}{2}(d-l)})$, we obtain that

$$\Big|\widehat{S_j}(m)\Big| \lesssim q^{-1-\frac{d+l}{2}} \sum_{r=0}^{d-l} |A_r(\chi, \psi)|,$$

where $A_r(\chi,\psi)$ is given by

$$\sum_{t\in\mathbb{F}_q^*}\chi(-tj)\psi^{-(d+r)}(t)\sum_{s_1,\dots,s_l\in\mathbb{F}_q^*}$$

$$\times\,\chi(s_1+\cdots+s_l+m_1^3t^{-1}s_1^{-1}+\cdots+m_l^3t^{-1}s_l^{-1})\psi(s_1)\cdots\psi(s_l).$$

We now apply the result of Adolphson and Sperber [1, Theorem 4.2, Corollary 4.3] to see that for all $r=0,1,\dots,d-l$,

$$|A_r(\chi,\psi)|\lesssim q^{\frac{l+1}{2}}.$$

This completes the proof. □

7. Proof of the second part of Theorem 3.1. As in the proof of the first part of Theorem 3.1, it suffices to prove the following estimation.

THEOREM 7.1. *Let $||x||_n=x_1^n+x_2^n$ for $x\in\mathbb{F}_q^2$ and $n\geq 2$. Suppose that q is a prime number and $j\neq 0$. Then if $m\neq(0,0)$, then*

$$\left|\widehat{S}_j(m)\right|=\left|q^{-2}\sum_{\{x\in\mathbb{F}_q^2:||x||_n=j\}}\chi(-x\cdot m)\right|\lesssim q^{-\frac{3}{2}},$$

and if $m=(0,0)$, then

$$\widehat{S}_j(m)=q^{-1}+O(q^{-\frac{3}{2}})\approx q^{-1}.$$

To prove Theorem 7.1, we observe that for $j\neq 0$ and $m\in\mathbb{F}_q^2$,

$$\widehat{S}_j(m)=q^{-2}\sum_{\{x\in\mathbb{F}_q^2:||x||_n=j\}}\chi(-x\cdot m)$$

$$=q^{-1}\delta(m)+q^{-3}\sum_x\sum_{t\in\mathbb{F}_q^*}\chi(t(||x||_n-j))\chi(-x\cdot m),$$

where $\delta(m)=1$ if $m=(0,0)$ and 0 otherwise.

First, we shall prove the second part of Theorem 7.1. Let ψ be a multiplicative character of $\mathbb{F}_q$ of order $h=gcd(n,q-1)$. For each $i=1,2,\dots,(h-1)$, we denote by β_i a nonnegative integer. Then by Theorem 6.5, we see that

$$\left(\sum_{s\in\mathbb{F}_q}\chi(ts^n)\right)^2$$

$$=\sum_{\beta_1+\cdots+\beta_{h-1}=2}\frac{2!}{\beta_1!\cdots\beta_{h-1}!}\psi^{-(\beta_1+\cdots+(h-1)\beta_{h-1})}(t)q^2\left(\widehat{\psi}(-1)\right)^{\beta_1}\cdots\left(\widehat{\psi^{h-1}}(-1)\right)^{\beta_{h-1}}.$$

It therefore follows that

$$\widehat{S}_j(0,0)=q^{-1}+\sum_{\beta_1+\cdots+\beta_{h-1}=2}\frac{2!}{\beta_1!\cdots\beta_{h-1}!}\widehat{\psi^{-\gamma(h,\beta)}}(j)\left(\widehat{\psi}(-1)\right)^{\beta_1}\cdots\left(\widehat{\psi^{h-1}}(-1)\right)^{\beta_{h-1}},$$

where $\gamma(h,\beta)$ is given by $\beta_1 + 2\beta_2 + \cdots + (h-1)\beta_{h-1}$.

Since $\widehat{\psi}(v) = O(q^{-\frac{1}{2}})$ for each multiplicative character ψ and $v \in \mathbb{F}_q^*$, we conclude that

$$\widehat{S}(0,0) = q^{-1} + O(q^{-\frac{3}{2}}) \approx q^{-1}.$$

This completes the proof of the second part of Theorem 7.1. ∎

It remains to prove the first part of Theorem 7.1. The cohomological interpretation can be used to estimate the exponential sums. We now introduce the cohomology theory based on the work of the authors in [4] and [1]. Let g be a polynomial given by

$$g = \sum_{\alpha \in J} A_\alpha x^\alpha \in \mathbb{F}_q[x_1, \ldots, x_d], \tag{7.1}$$

where J is a finite subset of $(\mathbb{N} \cup \{0\})^d$, and $A_\alpha \neq 0$ if $\alpha \in J$. We denote by $\sum(g)$ the Newton polyhedron of g which is the convex hull in $\mathbb{R}^d$ of the set $J \cup (0, \ldots, 0)$. For any face σ (of any dimension) of $\sum(g)$, we put

$$g_\sigma = \sum_{\alpha \in \sigma \cap J} A_\alpha x^\alpha.$$

DEFINITION 7.2. *Let $g \in \mathbb{F}_q[x_1, \ldots, x_d]$ be a polynomial as in* (7.1). *We say that g is nondegenerate with respect to $\sum(g)$ if for every face σ of $\sum(g)$ that does not contain the origin, the polynomials*

$$\frac{\partial g_\sigma}{\partial x_1}, \ldots, \frac{\partial g_\sigma}{\partial x_d}$$

have no common zero in $(\bar{\mathbb{F}}_q^{\,*})^d$, where $\bar{\mathbb{F}}_q$ denotes an algebraic closure of $\mathbb{F}_q$. We say that g is commode with respect to $\sum(g)$ if for each $k = 1, 2, \ldots, d$, g contains a term $A_k x_k^{\alpha_k}$ for some $\alpha_k > 0$ and $A_k \neq 0$.

The general version of the following theorem can be found in [4, Theorem 9.2].

THEOREM 7.3. *Let q be a prime number. Suppose that $g : \mathbb{F}_q^d \to \mathbb{F}_q, d \geq 2$, is commode and nondegenerate with respect to $\sum(g)$. Then*

$$\sum_{x \in \mathbb{F}_q^d} \chi(g(x)) = O(q^{\frac{d}{2}}).$$

Proof. We now prove the first part of Theorem 7.1. Since $m \neq (0,0)$, we have $\sum_{x \in \mathbb{F}_q^2} \chi(-x \cdot m) = 0$. We therefore see that for $j \neq 0$,

$$\widehat{S_j}(m) = q^{-3} \sum_{(t,x_1,x_2) \in \mathbb{F}_q^* \times \mathbb{F}_q^2} \chi(g(t,x_1,x_2)) = q^{-3} \sum_{(t,x_1,x_2) \in \mathbb{F}_q^3} \chi(g(t,x_1,x_2)),$$

where $g(t,x_1,x_2) = tx_1^n + tx_2^n - m_1x_1 - m_2x_2 - jt$.

If $m_1 \cdot m_2 \neq 0$, then g is commode. By Theorem 7.3, it suffices to show that g is nondegenerate with respect to $\sum(g)$. Note that $\sum(g)$ has five zero-dimensional faces, eight one-dimensional faces, and three two-dimensional faces which do not contain the origin. It is easy to show that for every face σ of $\sum(g)$ that does not contain the origin, the polynomials

$$\frac{\partial g_\sigma}{\partial t}, \frac{\partial g_\sigma}{\partial x_1}, \frac{\partial g_\sigma}{\partial x_2}$$

have no common zero in $(\mathbb{F}_q^*)^3$ because we may assume that q is sufficiently large and so n is not congruent to 0 modulo q. This implies that g is nondegenerate with respect to $\sum(g)$. We now assume that $m_1 \cdot m_2 = 0$. Without loss of generality, we may assume that $m_1 \neq 0$, and $m_2 = 0$ because $m \neq (0,0)$. By using Theorem 6.5, we obtain that for a multiplicative character ψ of $\mathbb{F}_q$ of order $h = gcd(n, q-1)$,

$$\widehat{S_j}(m) = q^{-3} \sum_{(t,x_1)\in \mathbb{F}_q^* \times \mathbb{F}_q} \chi(tx_1^n - m_1 x_1 - jt) \sum_{k=1}^{h-1} \psi^{-k}(t) q \widehat{\psi^k}(-1)$$

$$= q^{-2} \sum_{k=1}^{h-1} \widehat{\psi^k}(-1) \sum_{(t,x_1)\in \mathbb{F}_q^* \times \mathbb{F}_q} \psi^{-k}(t)\chi(tx_1^n - m_1 x_1 - jt)$$

$$\lesssim q^{-2} q^{-\frac{1}{2}} \sum_{k=1}^{h-1} |R_k(\psi^{-k}, \chi)|,$$

where $R_k(\psi^{-k}, \chi)$ is given by

$$\sum_{(t,x_1)\in \mathbb{F}_q^* \times \mathbb{F}_q} \psi^{-k}(t)\chi(tx_1^n - m_1 x_1 - jt).$$

For each $k = 1, 2, \ldots, h-1$, define $\psi^{-k}(0) = 0$. Then we can obtain that

$$R_k(\psi^{-k}, \chi) = \sum_{(t,x_1)\in \mathbb{F}_q \times \mathbb{F}_q} \psi^{-k}(t)\chi(tx_1^n - m_1 x_1 - jt).$$

Applying Theorem 7.3, we have

$$R_k(\psi^{-k}, \chi) = O(q).$$

This completes the proof. □

8. Proof of Theorem 3.4 and Corollary 3.6. As we mentioned in the introduction, this is a simple variation on the proof of Theorem 3.1. Indeed,

$$\#\{(x,y) \in E \times F : ||x-y||_n = j\}$$

$$= q^{2d} \sum_m \overline{\widehat{E}(m)} \widehat{F}(m) \widehat{S}_j(m)$$

$$= \#E \cdot \#F \cdot \widehat{S}_j(0, \ldots, 0) + q^{2d} \sum_{m\neq(0,\ldots,0)} \overline{\widehat{E}(m)} \widehat{F}(m) \widehat{S}_j(m) = \mathrm{I} + \mathrm{II}.$$

By the second part of Theorem 5.1 (or Theorem 7.1),

$$\mathrm{I} \lesssim \#E \cdot \#F \cdot q^{-1}.$$

Applying Cauchy–Schwarz, Theorem 5.1 (or Theorem 7.1), and Plancherel, we see that

$$|\mathrm{II}| \lesssim q^{2d} q^{-\frac{d+1}{2}} \sum_{m\neq(0,\ldots,0)} |\widehat{E}(m)||\widehat{F}(m)|$$

$$\leq q^{2d} q^{-\frac{d+1}{2}} \left(\sum_m |\widehat{E}(m)|^2 \right)^{\frac{1}{2}} \cdot \left(\sum_m |\widehat{F}(m)|^2 \right)^{\frac{1}{2}}$$

$$\leq q^{2d} q^{-\frac{d+1}{2}} q^{-d} \left(\sum_x |E(x)|^2 \right)^{\frac{1}{2}} \cdot \left(\sum_x |F(x)|^2 \right)^{\frac{1}{2}}$$

$$= q^{\frac{d-1}{2}} \cdot \sqrt{\#E} \cdot \sqrt{\#F}.$$

This completes the proof of Theorem 3.4. □

Proof. In order to prove Corollary 3.6, we observe that by the second part of Theorem 5.1 (or Theorem 7.1),

$$\mathrm{I} \gtrsim \#E \cdot \#F \cdot q^{-1}.$$

On the other hand, we have seen above that

$$|\mathrm{II}| \lesssim q^{\frac{d-1}{2}} \cdot \sqrt{\#E} \cdot \sqrt{\#F},$$

and the result follows by a direct comparison. □

REFERENCES

[1] A. Adolphson and S. Sperber, *Exponential sums and Newton polyhedra: Cohomology and estimates*, Ann. of Math. (2), 130 (1989), pp. 367–406.

[2] P. Agarwal and J. Pach, *Combinatorial Geometry*, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley and Sons, New York, 1995.

[3] J. Bourgain, N. Katz, and T. Tao, *A sum-product estimate in finite fields, and applications*, Geom. Funct. Anal., 14 (2004), pp. 27–57.

[4] J. Denef and F. Loeser, *Weights of exponential sums, intersection cohomology, and Newton polyhedra*, Invent. Math., 106 (1991), pp. 275–294.

[5] W. Duke and H. Iwaniec, *A relation between cubic exponential and Kloosterman sums*, in *A Tribute to Emil Grossweld: Number Theory and Related Analysis*, Contemp. Math. 143, AMS, Providence, RI, 1993, pp. 255–258.

[6] P. Erdös, *On sets of distances of n points*, Amer. Math. Monthly, 53 (1946), pp. 248–250.

[7] K. J. Falconer, *On the Hausdorff dimensions of distance sets*, Mathematika, 32 (1985), pp. 206–212.

[8] B. J. Green, *Restriction and Kakeya Phenomena*, Lecture notes, 2003; available online at http://www.dpmms.cam.ac.uk/ bjg23/rkp.html.

[9] A. Iosevich and M. Rudnev, *Erdös distance problem in vector spaces over finite fields*, Trans. Amer. Math. Soc., 359 (2007), pp. 6127–6142.

[10] H. Iwaniec and E. Kowalski, *Analytic Number Theory*, Amer. Math. Soc. Colloq. Publ. 53, AMS, Providence, RI, 2004.

[11] N. Katz, *Gauss sums, Kloosterman Sums, and Monodromy Groups*, Ann. of Math. Stud. 116, Princeton University Press Princeton, NJ, 1988.

[12] M. Lacey and W. McClain, *On an argument of Shkredov in the finite field setting*, Analytic Online Journal of Combinatorics, 2007; available online at http://www.ojac.org.

[13] R. Lidl and H. Niederreiter, *Finite Fields*, Cambridge University Press, Cambridge, UK, 1997.

[14] J. Matoušek, *Lectures on Discrete Geometry*, Grad. Texts in Math. 212, Springer-Verlag, New York, 2002.

[15] G. Mockenhaupt and T. Tao, *Restriction and Kakeya phenomena for finite fields*, Duke Math. J., 121 (2004), pp. 35–74.

[16] H. Niederreiter, *The distribution of values of Kloosterman sums*, Arch. Math., 56 (1991), pp. 270–277.

[17] I. Shparlinski, *On the set of distances between two sets over finite fields*, Int. J. Math. Math. Sci., (2006), pp. 1–5.
[18] E. Stein and R. Shakarchi, *Fourier Analysis*, Princeton Lect. Anal. 1, Princeton University Press, Princeton, NJ, 2003.
[19] T. Tao, *Finite Field Analogues of Erdös, Falconer, and Furstenberg Problems*, preprint.
[20] A. Weil, *On some exponential sums*, Proc. Nat. Acad. Sci. U.S.A., 34 (1948), pp. 204–207.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 136–154

INTEGER EXACT NETWORK SYNTHESIS PROBLEM*

S. N. KABADI†, J. YAN‡, D. DU†, AND K. P. K. NAIR†

Abstract. Given an integer, nonnegative, symmetric matrix $R = (r_{ij})_{n\times n}$, we consider the problem of synthesizing an undirected network G on node set $V = \{1, 2, \ldots, n\}$ with nonnegative, integer edge capacities such that (i) for any pair $\{i, j\}$ of distinct nodes in V, the value of maximum flow between i and j in G equals exactly r_{ij} and (ii) the sum of capacities of edges in G is minimum. Chou and Frank [*IEEE Trans. Circuit Theory*, CT-17 (1970), pp. 192–197] claim to give an algorithm for this problem. But, Schrijver [*Algorithms Combin.* 24, Springer-Verlag, Berlin, 2003, pp. 1049–1057] gives a counter-example to their claim. We present an $O(n^2)$ algorithm for the problem.

Key words. combinatorial algorithm, cut-tree, network flows, strongly polynomial algorithm

AMS subject classifications. 90C27, 90B10

DOI. 10.1137/050641776

1. Introduction.

Let $R = (r_{ij})_{n\times n}$ be a symmetric, nonnegative matrix of minimum flow requirements between all pairs of distinct nodes in the set $V = \{1, 2, \ldots, n\}$, where $r_{ii} = 0$ $(i = 1, 2, \ldots, n)$. We call an undirected network $G = [V, E, u]$ on node set V with edge set E and nonnegative edge capacities $\{u_e : e \in E\}$ a *realization* of R if and only if for every pair $\{i, j\}$ of distinct nodes in V, the value of maximum flow between i and j in G is at least r_{ij}. We say that G is an *exact realization* of R if and only if for every pair $\{i, j\}$ of distinct nodes in V, the value of maximum flow between i and j in G equals exactly r_{ij}. If R has an exact realization, then we say that it is *exactly realizable*.

The *network synthesis problem* (NSP) constructs a realization of R with a minimum sum of edge capacities, while the *exact network synthesis problem* (ENSP) constructs an exact realization of R with a minimum sum of edge capacities, or else concludes that it is not exactly realizable. In these two problems, if the elements of R are integers and we require all of the edge capacities of G to be integers, then we get, respectively, *the integer network synthesis problem* (INSP) and *the integer exact network synthesis problem* (IENSP).

The main focus of this work is the last problem and our main contribution is to present an $O(n^2)$ combinatorial algorithm for the IENSP. We review existing results for these four problems below.

The NSP (and its generalization to the case of synthesizing a network with minimum weighted sum of edge capacities) has a polynomial size linear programming formulation [8] which can be solved in strongly polynomial time using Tardos' algorithm [26]. However, the Tardos' algorithm is not very efficient in practice, nor does it provide any insight into the combinatorial structure of the problem. In [9, 21] effi-

*Received by the editors October 3, 2005; accepted for publication (in revised form) June 30, 2008; published electronically November 14, 2008.
http://www.siam.org/journals/sidma/23-1/64177.html

†Faculty of Business Administration, University of New Brunswick, Fredericton, NB, E3B 5A3, Canada (kabadi@unb.ca, ddu@unb.ca, nairk@unb.ca). The research of these authors was supported in part by research grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

‡Faculty of Business Administration, University of New Brunswick, Fredericton, NB, E3B 5A3, Canada (jyan@unb.ca), and School of Mathematics and Systems Sciences, Shandong University, Jinan 250100, People's Republic of China (yanj@sdu.edu.cn).

cient, combinatorial, and strongly polynomial algorithms are presented for the NSP. The Gomory–Hu algorithm in [9] has a computational complexity of $O(n^2)$. Also, when all of the elements of the matrix R are integers, then the edge capacities in the final network are multiples of half. Alternately, combinatorial algorithms for the NSP are also presented in [10, 25].

In [3] and independently in [24], combinatorial algorithms of computational complexity $O(n^2)$ are presented for the INSP and it is shown that whenever $\max\{r_{ij} : j \in V - \{i\}\} > 1$ for all $i \in V$, the problem has *integer rounding property* (i.e., the difference between the sum of edge capacities in the optimal networks for the integer and continuous versions of the problem is less than 1). Alternate algorithms for the problem are given in [18, 23]. In [6], a strongly polynomial algorithm is given for a generalization of this problem to one in which we want to increase given integer edge capacities by integer amounts so as to obtain a realization of a given integer matrix R such that the sum of additional edge capacities is minimum.

For the weighted cases of the NSP and INSP, strongly polynomial combinatorial algorithms are known only for the special case in which the network is restricted to be a cycle [12, 13]. Results on generalizations of these problems are reported in [11, 14] for the case of 2-commodity flows, in [7, 17] for the case of hop-constrained flows, and in [1, 2, 16] for the case of multipath flows [19].

As pointed out in [23], a modification of the Gomory–Hu algorithm in [9] produces an optimal solution to the NSP that is also an optimal solution to the corresponding ENSP whenever the latter has a feasible solution. It is shown in [15] that generalization of the ENSP to the case of 2-path flows [19] is NP-hard.

In [3], Chou and Frank claim to give an algorithm for the IENSP. However, in [23], Schrijver gives a counterexample to their claim, and hence leaves open the status of the problem.

The purpose of this work is to devise an $O(n^2)$ combinatorial algorithm for the IENSP. To facilitate presentation of our algorithm, we define four subproblems and present an algorithm for each of them. The final algorithm for the IENSP invokes these algorithms as subroutines.

The rest of this paper is organized as follows. After presenting notation and some basic results in section 2, we discuss the four subproblems in sections 3, 4, 5, and 6, respectively. Our algorithm for the IENSP is then presented in section 7.

2. Notation and preliminaries. Throughout this paper, all networks are simple, undirected, and edge-capacitated, and all of the edge capacities considered are nonnegative. Let $G = [V, E, u]$ be a network on node set $V = \{1, 2, \ldots, n\}$ with edge set E and edge capacities $\{u_e : e \in E\}$.

The *degree* $\deg_G(v)$ of a node $v \in V$ is the number of edges incident to it in G. For any nonempty set $X \subseteq V$, we denote the sum of capacities of all of the edges in the subgraph of G induced by node set X as $u[X] = \sum\{u_e : e = (i, j) \in E,\ \{i, j\} \subseteq X\}$. For any nonempty and proper subset $X \subset V$ and its *complement* $\bar{X} = V - X$, we denote the *cut* separating node sets X and $\bar{X}$ as $[X, \bar{X}] = \{e : e = (i, j) \in E,\ i \in X,\ j \in \bar{X}\}$ and denote the *capacity* of cut $[X, \bar{X}]$ as $\delta_u[X] = \sum\{u_e : e \in [X, \bar{X}]\}$.

DEFINITION 2.1. *Two networks $G^1 = [V, E^1, u^1]$ and $G^2 = [V, E^2, u^2]$ on the same node set V and with edge capacities $\{u^1_e : e \in E^1\}$ and $\{u^2_e : e \in E^2\}$ are said to be flow-equivalent if for any pair $\{i, j\}$ of distinct nodes in V, the maximum flow values between i and j in the two networks are the same. An edge-capacitated tree $T = [V, E^T, u^T]$ that is flow-equivalent to a network G is called a flow-tree of G.*

The following result is easy to prove and implicit in [9, 21].

LEMMA 2.2. *Let $T^1 = [V, E^1, u^1]$ and $T^2 = [V, E^2, u^2]$ be a pair of edge-capacitated trees on the same node set V. For each $i \in \{1, 2\}$, let $u^i_{\min} = \min\{u^i_e : e \in E^i\}$, and let $T^i_1, T^i_2, \ldots, T^i_{\ell_i}$ be the subtrees formed by deleting from T^i all the edges of capacity $u^i_{\min}$. Then T^1 and T^2 are flow-equivalent if and only if $u^1_{\min} = u^2_{\min}$, $\ell_1 = \ell_2 = \ell$ and there exists a permutation ϕ on $\{1, 2, \ldots, \ell\}$ such that T^1_j is flow-equivalent to $T^2_{\phi(j)}$ for each $j \in \{1, 2, \ldots, \ell\}$.*

DEFINITION 2.3. *Given a tree $T = [V, E]$ and any edge $f = (i, j) \in E$, let V_{fi} and V_{fj} be the node sets of the two subtrees formed by deleting edge f from T such that $i \in V_{fi}$ and $j \in V_{fj}$. Then the cut $[V_{fi}, V_{fj}]$ is the fundamental cut of T corresponding to edge f.*

DEFINITION 2.4 (see [5]). *Given a network $G = [V, E, u]$, an edge-capacitated tree $T = [V, E^T, u^T]$ is said to be a cut-tree of G if and only if for each edge $f = (i, j) \in E^T$, the fundamental cut $[V_{fi}, V_{fj}]$ of T corresponding to edge f is a minimum capacity cut separating nodes i and j in G, and $u^T_f = \delta_u[V_{fi}]$.*

We also need the following results.

LEMMA 2.5 (see [9]). *Let $T = [V, E^T, u^T]$ be a cut-tree of a network $G = [V, E, u]$. Then T is a flow-tree of G. Moreover, for any pair of distinct nodes $\{x, y\}$ in V, let $f = (i, j) \in E^T$ be an edge on the unique path in T joining nodes x and y with the smallest value of u^T_f. Then the fundamental cut $[V_{fi}, V_{fj}]$ of T corresponding to edge f is a minimum capacity cut separating nodes x and y in G.*

THEOREM 2.6 (see [9]). *For a symmetric, $n \times n$, nonnegative matrix R, let G^R be a complete network on node set $V = \{1, 2, \ldots, n\}$ with the capacity of each edge (i, j) equal to r_{ij}. Then we have the following.*

1. *Every maximum weight spanning tree $T = [V, E]$ of G^R with edge capacities $u_e = r_{ij}$ for all $e = (i, j) \in E$ is a realization of R.*
2. *The following three statements are equivalent.*
   (a) *R is exactly realizable.*
   (b) *$r_{ij} \geq \min\{r_{ik}, r_{kj}\}$ for all distinct $i, j, k \in V$.*
   (c) *Every maximum weight spanning tree $T = [V, E]$ of G^R with edge capacities $u_e = r_{ij}$ for all $e = (i, j) \in E$ is an exact realization of R.*

The following is a corollary to the above.

COROLLARY 2.7 (see [9, 21]). *There exists an $O(n^2)$ algorithm to test whether a given matrix is exactly realizable.*

LEMMA 2.8 (see [3, 21]). *If R is exactly realizable, then there exists a maximum weight spanning tree of G^R that is a path.*

We give below a version of the Gomory–Hu algorithm [9] for the NSP that we will require in later sections.

Algorithm GOMORY–HU

Input. An $n \times n$, symmetric, nonnegative matrix R.

Output. An optimal solution $G^* = [V, E^*, u^*]$ to the instance of the NSP.

Step 0. Select a Hamiltonian cycle $\mathcal{H} = v_1 v_2 \cdots v_n v_1$ on node set V. Find a maximum weight spanning tree $T = [V, E]$ in G^R. Set $E^0 = E$, $F^0 = [V, E^0]$, $r^0_e = r_{ab}$, and $u^0_e = 0$ for all $e = (a, b)$, where $a, b \in V$ and $a \neq b$. Initialize $i = 0$.

Step 1. Arbitrarily choose a connected component $T^i = [V^i, \hat{E}^i]$ of F^i with at least two nodes. Let $V^i = \{\ell_1, \ell_2, \ldots, \ell_{m_i}\}$, where $\ell_1, \ell_2, \ldots, \ell_{m_i}$ appears in that order along the cycle $\mathcal{H}$. Define cycle $\mathcal{C}^i = \ell_1 \ell_2 \cdots \ell_{m_i} \ell_1$. Set

$$\theta^i = \min\{r^i_e : e \in \hat{E}^i\};$$

$$\Delta^i u_e = \begin{cases} \frac{1}{2}\theta^i & \text{if } e \text{ is an edge in } \mathcal{C}^i, \\ 0 & \text{otherwise;} \end{cases}$$
$$u_e^{i+1} = u_e^i + \Delta^i u_e \ \forall e;$$
$$r_e^{i+1} = \begin{cases} r_e^i - \theta^i & \forall\ e \in \hat{E}^i, \\ r_e^i & \forall\ e \in E^i - \hat{E}^i. \end{cases}$$

Step 2. Delete from E^i all the edges with $r_e^{i+1} = 0$ to get edge set E^{i+1}. Set

$$F^{i+1} = [V, E^{i+1}],$$
$$i = i + 1.$$

1. If F^i contains only isolated nodes, then set

$$E^* = \{e : u_e^i > 0\},$$
$$u_e^* = u_e^i \ \forall e \in E^*.$$

   Output $G^* = [V, E^*, u^*]$ and stop.
2. Else, go to Step 1.

THEOREM 2.9 (see [9]). *Let $\pi_i = \max\{r_{ij} : j \in V - \{i\}\}$ for each $i \in V$. The network G^* output by Algorithm* GOMORY–HU *is an optimal solution to the NSP with $u^*[V] = \frac{1}{2}\sum_{i \in V} \pi_i$. Also, if r_e^0 is even for all $e \in T$ (a maximum weight spanning tree in G^R), then u_e^* is integer for all e.*

THEOREM 2.10 (see [23]). *Suppose R is exactly realizable. If in Step 0 of Algorithm* GOMORY–HU *we choose a maximum weight spanning tree in G^R that is a path and choose $\mathcal{H}$ as the cycle obtained by joining the endnodes of the path, then the network G^* output by Algorithm* GOMORY–HU *is an optimal solution to the ENSP on R.*

The following result generalizes Theorem 2.10.

THEOREM 2.11. *Suppose R is exactly realizable. Let T be any maximum weight spanning tree in G^R. Let the Hamiltonian cycle $\mathcal{H}$ in Step 0 of Algorithm* GOMORY–HU *be such that every fundamental cut of T contains precisely two edges of $\mathcal{H}$. Then the network G^* output by Algorithm* GOMORY–HU *has T as its cut-tree, and hence it is an optimal solution to the ENSP on R.*

The proof of this theorem follows from the proof of Theorem 6.5. Hence, we shall not give details here. The following simple scheme for constructing such a Hamiltonian cycle was pointed out to us by Punnen [22]. For any planar embedding of T, join the leaf nodes of T to obtain a Halin graph G [4]. Then a Hamiltonian cycle in G satisfies the desired property and can be obtained in linear time [4].

We explain the key ideas of our approach and the organization of the rest of this paper. To avoid technical complication, we assume that $r_{ij} > 1$ for all i, j. The cases when some elements of R are zero or one are easy to handle, as we show in section 7. First, we check (using Theorem 2.6) whether matrix R is exactly realizable. If the answer is positive, then we use the observations below that follow from the previous discussion. Let M^R be the set of all the maximum weight spanning trees of G^R. Then (i) M^R is the set of all cut-trees of all the possible exact realizations of R and (ii) for any $T \in M^R$, M^R is the set of all the trees that are flow-equivalent to T.

We consider a slight modification of the IENSP which we call the *optimal cut-tree realization problem* (OCRP): Given an edge-capacitated tree $T = [V, E, u]$ with integer edge capacity $u_e > 1$ for all $e \in E$, construct a network $G^* = [V, E^*, u^*]$ that

has T as its cut-tree such that (i) u_e^* is an integer for each $e \in E^*$ and (ii) $u^*[V] = \sum_{e \in E^*} u_e^*$ minimum.

We present an $O(n^2)$ algorithm for this problem in section 6 and show that the optimal objective function value equals

$$\frac{\sum_{i \in V} \pi_i + |V_1^e(T)| + |V_2^o(T)|}{2},$$

where for each node $i \in V$, $\pi_i = \max\{u_e : e = (i,j) \in E\}$, $V_1 = \{i \in V : \pi_i \text{ is odd}\}$, $V_2 = V - V_1$, and $V_1^e(T)$ ($V_2^o(T)$) is the set of nodes in V_1 (V_2) each of which has an even (odd) number of odd edges (edges with capacity u_e odd) incident to it in T.

The values $\{\pi_i : i \in V\}$ are the same for each $T \in M^R$. Hence, if we choose $T \in M^R$ with a minimum value of $|V_1^e(T)| + |V_2^o(T)|$, then an optimal solution to the corresponding instance of the OCRP will be an optimal solution to the IENSP. We therefore consider the problem of finding such a tree, which we call the *flow-equivalent tree problem* (FETP) and present an $O(n^2)$ algorithm for it in section 5.

The algorithms for the OCRP and the FETP require as subroutines algorithms for two other problems, which we call the *tree path covering problem* (TPCP) and the *degree constrained spanning forest problem* (DCSFP), respectively. We therefore first discuss these two problems and present efficient algorithms for them, respectively, in sections 3 and 4.

Finally in section 7, we combine the results in the previous sections to present our algorithm for the IENSP.

3. Subproblem 1: The tree path covering problem (TPCP).

Statement of the problem. We are given a tree $T = [V, E]$ and partitions (V_1, V_2) and (E_1, E_2) of node set V and edge set E into sets of *odd-even* nodes and *odd-even* edges, respectively, such that each node in V_1 has at least one edge in E_1 incident to it. The problem is to find a set $\mathcal{P} = \{p_1, p_2, \ldots, p_k\}$ of edge-disjoint paths in T covering all of the edges in E_1 and not containing any edge in E_2 such that (i) every node in V_1 (i.e., every odd node) is an endnode of some path in $\mathcal{P}$ and (ii) k is minimum. We denote the minimum value of k by $\lambda(T)$.

Let $V_1^o(T) \subseteq V_1$ and $V_2^o(T) \subseteq V_2$ be such that each node in $V_1^o(T) \cup V_2^o(T)$ has an odd number of odd edges (edges in E_1) incident to it in T. Let $V_1^e(T) = V_1 - V_1^o(T)$ and $V_2^e(T) = V_2 - V_2^o(T)$. Then each node in $V_1^e(T) \cup V_2^e(T)$ has an even number of odd edges incident to it in T. The following observation is easy to verify.

OBSERVATION 3.1. *In any feasible solution to an instance of the TPCP, every node in $V_1^o(T) \cup V_2^o(T)$ must appear as an endnode of a path an odd number of times and every node in $V_1^e(T) \cup V_2^e(T)$ must appear as an endnode of a path an even number of times.*

THEOREM 3.2. *For any instance of the TPCP, an optimal set $\mathcal{P}^*$ of edge-disjoint paths with optimal value*

$$\lambda(T) = |\mathcal{P}^*| = \frac{|V_1| + |V_1^e(T)| + |V_2^o(T)|}{2}$$

can be computed in $O(n)$ time.

Proof. From the statement of the problem and Observation 3.1, we infer that in any feasible solution, each node in $V_1^e(T)$ occurs as an endnode of some path at least

twice and each node in $V_2^o(T)$ occurs as an endnode at least once. Thus,

$$\begin{aligned}\lambda(T) &= \frac{\text{total number of endnodes in an optimal solution}}{2} \\ &\geq \frac{|V_1| + |V_1^e(T)| + |V_2^o(T)|}{2}.\end{aligned}$$

Let X be the multiset containing nodes in $V_1^o(T)$, $V_2^o(T)$, and two copies of each node in $V_1^e(T)$. The cardinality of X is obviously even. Let E^M be a perfect matching of elements of X such that no element of $V_1^e(T)$ is matched with the copy of itself. Then each connected component of $G = [V, E_1 \cup E^M]$ (here the duplicated nodes are contracted to recover the original node set V) is an Eulerian graph [20]. We can find Eulerian tours $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_\ell$ in the connected components of G in $O(|E_1 \cup E^M|) = O(n)$ time [20]. It is easy to see that since each node in V_1 has at least one edge of E_1 incident to it, such tours can be chosen such that no two edges in E^M are adjacent in any tour. Deleting the edges in E^M from these Eulerian tours gives us a set of paths $\mathcal{P}^*$ with

$$|\mathcal{P}^*| = |E^M| = \frac{|V_1| + |V_1^e(T)| + |V_2^o(T)|}{2}.$$

This proves the theorem. □

4. Subproblem 2: The degree constrained spanning forest problem (DCSFP) on a complete graph.

Statement of the problem. Given an integer $n \geq 2$ and positive integer weights $w_1, w_2, \ldots, w_n$, find a spanning forest $F = [V, E]$ on node set $V = \{1, 2, \ldots, n\}$ with a maximum number of edges $|E|$ such that $\deg_F(i) \leq w_i$ for all $i \in V$.

THEOREM 4.1. *For any instance of the DCSFP, an optimal spanning forest $F^* = [V, E^*]$ with optimal value*

$$|E^*| = \min\left\{\left\lfloor \frac{\sum_{i=1}^n w_i}{2} \right\rfloor, n-1\right\}$$

can be computed in $O(n)$ time.

Proof. Each of $\lfloor \frac{\sum_{i=1}^n w_i}{2} \rfloor$ and $(n-1)$ is obviously an upper bound on the number of edges in any optimal solution to the DCSFP. Next, we construct a feasible spanning forest $F^* = [V, E^*]$ with the claimed value.

Partition the node set V into two sets $V' = \{i \in V : w_i \geq 2\}$ and $V'' = \{i \in V : w_i = 1\}$.

If $|V'| = 0$, then a maximum cardinality matching on node set V'' is the desired F^* with

$$|E^*| = \left\lfloor \frac{|V|}{2} \right\rfloor = \left\lfloor \frac{\sum_{i=1}^n w_i}{2} \right\rfloor = \min\left\{\left\lfloor \frac{\sum_{i=1}^n w_i}{2} \right\rfloor, n-1\right\}.$$

If $|V'| \geq 1 \geq |V''|$, then a path on node set V with the node in V'', if it exists, as an endnode of the path is our desired F^* with $|E^*| = n - 1$.

So in the following we assume $|V'| \geq 1$ and $|V''| \geq 2$. First, we form a path P that visits each node in V' and two distinct nodes $u, v \in V''$ exactly once such that u and v are endnodes of this path. Next, let $\widehat{V}' = \{i \in V' : w_i - 2 > 0\}$ and $\widehat{V}'' = V'' - \{u, v\}$. Partition $\widehat{V}''$ into two parts $\widehat{V}_1''$ and $\widehat{V}_2'' = \widehat{V}'' - \widehat{V}_1''$ such that

$|\widehat{V}_1''| = \min\left\{\sum_{i\in V'}(w_i - 2), |V''| - 2\right\}$. Then, arbitrarily join each node in $\widehat{V}_1''$ with some node in $\widehat{V}'$ to form a forest F (a set of stars) such that $\deg_F(i) = 1$ for all $i \in \widehat{V}_1''$ and $\deg_F(i) \leq w_i - 2$ for all $i \in \widehat{V}'$. Finally, find a maximum cardinality matching M in $\widehat{V}_2''$. We now show that the superimposing of P, F, and M gives us the desired F^*. Evidently, the superimposed graph is a spanning forest and satisfies the degree constraints, and hence is feasible. Moreover, the claimed objective function value follows from

$$|E^*| = (|V'| + 1) + |\widehat{V}_1''| + \left\lfloor \frac{|\widehat{V}_2''|}{2} \right\rfloor = (|V'| + 1) + |\widehat{V}_1''| + \left\lfloor \frac{|V''| - 2 - |\widehat{V}_1''|}{2} \right\rfloor$$

$$= \left\lfloor \frac{2|V'| + |V''| + |\widehat{V}_1''|}{2} \right\rfloor = \left\lfloor \frac{2|V'| + |V''| + \min\left\{\sum\limits_{i\in V'}(w_i - 2), |V''| - 2\right\}}{2} \right\rfloor$$

$$= \min\left\{ \left\lfloor \frac{\sum_{i=1}^{n} w_i}{2} \right\rfloor, n - 1 \right\}.$$

All the operations involved can be done in $O(n)$ time. This proves the theorem. □

5. Subproblem 3: The flow-equivalent tree problem (FETP). We first introduce notation useful in this section. For any edge-capacitated tree $T = [V, E, u]$ with integer-valued edge capacities, let $\pi_i = \max\{u_e : e = (i, j) \in E\}$ for all $i \in V$. We call $V_1 = \{i \in V : \pi_i \text{ is odd}\}$ and $V_2 = V - V_1$ the sets of *odd* and *even* nodes, respectively. We call $E_1 = \{e \in E : u_e \text{ is odd}\}$ and $E_2 = E - E_1$ the sets of *odd* and *even* edges of T, respectively. As defined in the previous section, let $V_1^e(T)$ ($V_2^o(T)$) be the set of nodes in V_1 (V_2) each of which has an even (odd) number of odd edges of T incident to it. It follows from Lemma 2.2 that the values $\{\pi_i : i \in V\}$ and therefore the node sets V_1 and V_2 are the same for all of the trees flow-equivalent to T.

Statement of the problem. Find an edge-capacitated tree $T^* = [V, E^*, u^*]$ that is flow-equivalent to a given edge-capacitated tree $T = [V, E, u]$ with integer-valued edge capacities such that the optimal objective function value $\lambda(T^*)$ of the corresponding instance of the TPCP with the node and edge partitions (V_1, V_2) and (E_1^*, E_2^*), respectively, is minimum.

We propose the following recursive algorithm for the FETP. The algorithm deletes from the given tree T all of the minimum capacity edges, recursively obtains optimal solutions (flow-equivalent trees) to the instances of the FETP on each of the subtrees obtained, and then optimally links these optimal subtrees to get an optimal solution to the given instance of the FETP.

Algorithm TREE-FINDING

Input: An edge-capacitated tree $T = [V, E, u]$ on node set $V = \{1, 2, \ldots, n\}$ with integer-valued edge capacities.

Output: A tree $T^* = [V, E^*, u^*]$ that is an optimal solution to the instance of the FETP.

Step 1. If $|V| \leq 2$, then output $T^* = T$ and stop. Else, compute $\{\pi_i : i \in V\}$, node partition (V_1, V_2), and edge partition (E_1, E_2). Find $u_{\min} = \min\{u_e : e \in E\}$. Let $T^i = (V^i, E^i)$, $i = 1, \ldots, \ell$, be the subtrees resulting from deletion of all of the edges of capacity $u_{\min}$ from tree T.

Step 2. For each $i \in \{1, 2, \ldots, \ell\}$, recursively find an optimal solution $T^{*i} = [V^i, E^{*i}, u^{*i}]$ to the FETP with tree T^i as input.

1. If $u_{\min}$ is even, then arbitrarily choose a node $x_i \in V^i$ for each $i \in \{1, 2, \ldots, \ell\}$. Let $\widetilde{E} = \{(x_i, x_{i+1}) : i \in \{1, 2, \ldots, \ell - 1\}\}$.
2. If $u_{\min}$ is odd, then set $V_1^i = V^i \cap V_1$, $V_2^i = V^i \cap V_2$, and $w_i = |V_1^{i,e}| + |V_2^{i,o}|$ for each $i \in \{1, 2, \ldots, \ell\}$, where $V_1^{i,e}$ is the set of nodes in V_1^i, each of which has an even number of odd edges of T^{*i} incident to it, and $V_2^{i,o}$ is the set of nodes in V_2^i, each of which has an odd number of odd edges of T^{*i} incident to it. Renumber the trees $T^{*1}, T^{*2}, \ldots, T^{*\ell}$, if necessary, such that for some integer m, $w_i \geq 1$ for any $i \leq m$ and $w_i = 0$ for any $i > m$.
   (a) If $m = 0$, then arbitrarily choose a node $x_i \in V^i$ for each $i \in \{1, 2, \ldots, \ell\}$. Let $\widetilde{E} = \{(x_i, x_{i+1}) : i \in \{1, 2, \ldots, \ell - 1\}\}$.
   (b) If $m = 1$, then choose $x_1 \in V_1^{1,e} \cup V_2^{1,o}$ and arbitrarily choose a node $x_i \in V^i$ for each $i \in \{2, \ldots, \ell\}$. Let $\widetilde{E} = \{(x_i, x_{i+1}) : i \in \{1, 2, \ldots, \ell - 1\}\}$.
   (c) If $m > 1$, then find an optimal solution $\bar{F} = [M, \bar{E}]$ on node set $M = \{1, 2, \ldots, m\}$ to the DCSFP with weights $w_1, w_2, \ldots, w_m$ as input. Define set $\widetilde{E}$ as follows:
   - For each edge $(i, j) \in \bar{E}$, associate with the edge distinct nodes $x \in V_1^{ie} \cup V_2^{io}$ and $y \in V_1^{je} \cup V_2^{jo}$ (i.e., no node is associated with two edges in $\bar{E}$ in this process) and add edge (x, y) to $\widetilde{E}$.
   - Arbitrarily delete one edge (a, b) from $\widetilde{E}$. Let $\hat{E} = (\cup_{i=1}^{\ell} E^{*i}) \cup \widetilde{E}$.
   - Let the node sets of the connected components of the graph $\hat{G} = [V, \hat{E}]$ be $\hat{V}^1, \hat{V}^2, \ldots, \hat{V}^q$. Without loss of generality, let $a \in \hat{V}^1$ and $b \in \hat{V}^q$. Arbitrarily choose a node $x_i \in \hat{V}^i$ for each $i \in \{2, \ldots, q - 1\}$. Add to $\widetilde{E}$ edges $\{a, x_2), (x_2, x_3), \ldots, (x_{q-1}, b)\}$.

Step 3. Let $E^* = (\cup_{i=1}^{\ell} E^{*i}) \cup \widetilde{E}$. Define

$$u_e^* = \begin{cases} u_e^{*i} & \text{if } e \in E^{*i} \text{ for some } i \in \{1, 2, \ldots, \ell\}, \\ u_{\min} & \text{if } e \in \widetilde{E}. \end{cases}$$

Output tree $T^* = [V, E^*, u^*]$ and stop.

The following two properties are easy to verify and will be useful in the proof of validity of the algorithm.

LEMMA 5.1. *For any edge-capacitated tree T on node set V, $|V_1| + |V_1^e(T)| + |V_2^o(T)|$ is an even number. Hence, for any two flow-equivalent trees T^1 and T^2 on the node set V,*

$$(|V_1^e(T^1)| + |V_2^o(T^1)|) \equiv \big(|V_1^e(T^2)| + |V_2^o(T^2)|\big) \bmod 2.$$

LEMMA 5.2. *For any tree T on a node set V and any $S \subseteq V$, let $\beta_T^S = \sum_{i \in S} \deg_T(i)$. Then the set $\bar{S} = V - S$ contains at least $\beta_T^S - 2|S| + 2$ leaf nodes of T.*

Proof. This follows from the fact that $\beta_T^{\bar{S}} = 2(|V| - 1) - \beta_T^S = 2|\bar{S}| - (\beta_T^S - 2|S| + 2)$. ∎

We will now prove the validity of Algorithm TREE-FINDING.

THEOREM 5.3. *For any instance of the FETP, Algorithm* TREE-FINDING *constructs an optimal tree in $O(n^2)$ time.*

Proof. For $|V| \leq 2$, the tree T is obviously an optimal solution to the given instance of the FETP. It follows from Lemma 2.2 and the inductive hypothesis that

for any value of $|V|$, the tree T^* output by the algorithm is flow-equivalent to the given tree T. Let us now prove the optimality of T^*.

Suppose T^* is optimal for any $|V| \leq k$ for some $k \geq 2$. Let us consider the case $|V| = k+1$. Let $T^1, T^2, \ldots, T^\ell$ be the subtrees of T obtained by deleting from T all of the edges with capacity $= u_{\min}$ in Step 1 of Algorithm TREE-FINDING. It follows from the induction hypothesis that for each $i \in \{1, 2, \ldots, \ell\}$, tree T^{*i} obtained recursively by the algorithm with tree T^i as input is an optimal solution to the corresponding instance of the FETP. Also, for each i such that $|V^i| \geq 2$, (V_1^i, V_2^i) is precisely the odd-even node partition in T^i.

Suppose $u_{\min}$ is even. By Theorem 3.2, for any tree T, the optimal objective function value of the TPCP on T is independent of the structure of even edges of T. It follows from this and Lemma 2.2 that the tree T^* obtained by the algorithm by joining trees $T^{*1}, T^{*2}, \ldots, T^{*\ell}$ using $(\ell - 1)$ edges each of capacity $u_{\min}$ is an optimal solution to the given instance of the FETP.

Suppose $u_{\min}$ is odd. Let integer m be as defined in Step 2 of the algorithm and let $w = |V_1^e(T^*)| + |V_2^o(T^*)|$. Then $\lambda(T^*) = \frac{|V_1|+w}{2}$. It is easy to see that (i) if $m = 0$, then $w = 2$, (ii) if $m = 1$, then $w = w_1$, and (iii) if $m > 1$, then $w = \sum_{i=1}^m w_i - 2|\bar{E}|$, where $\bar{F} = [M, \bar{E}]$ is the optimal solution to the DCSFP obtained in Step 2 of the algorithm. By Theorem 4.1, $|\bar{E}| = \min\{\frac{1}{2}\sum_{i=1}^m w_i, m-1\}$. Thus, in general, for any $m \geq 0$, $w = \sum_{i=1}^m w_i - 2z$, where $z = \min\{\frac{1}{2}\sum_{i=1}^m w_i, m-1\}$.

If $z = \frac{1}{2}\sum_{i=1}^m w_i$, then $w = 0$ and $\lambda(T^*) = \frac{|V_1|}{2}$. By Theorem 3.2, in this case T^* is obviously optimal.

Suppose $z = m - 1$. Then

$$\lambda(T^*) = \frac{|V_1| + \sum_{i=1}^m w_i - 2m + 2}{2}.$$

Consider any other feasible solution $\bar{T}$ to the problem. By Lemma 2.2, if we delete from $\bar{T}$ all of the edges of capacity $u_{\min}$, then we get subtrees $\bar{T}^1, \bar{T}^2, \ldots, \bar{T}^\ell$ that are flow-equivalent to subtrees $T^1, T^2, \ldots, T^\ell$ of T. Let $\bar{V}_1^{ie}$ ($\bar{V}_2^{io}$) be the set of nodes in V_1^i (V_2^i) each of which has an even (odd) number of odd edges of $\bar{T}^i$ incident to it, and let $\bar{w}_i = |\bar{V}_1^{ie}| + |\bar{V}_2^{io}|$. By optimality of T^{*i}, we have $\bar{w}_i \geq w_i$. Let $\bar{\bar{T}}$ be the tree obtained from $\bar{T}$ by contracting each subtree $\bar{T}^i$ to a supernode α_i, $i = 1, \ldots, \ell$. Let $S = \{\alpha_1, \ldots, \alpha_m\}$, $\beta_{\bar{\bar{T}}}^S = \sum_{i \in S} \deg_{\bar{\bar{T}}}(i)$, and let $X \subseteq \{\alpha_1, \ldots, \alpha_\ell\} - S$ be the nodes not in S that are leaf nodes of $\bar{\bar{T}}$. Then node set $\cup_{i=1}^m V^i$ contains at least $\sum_{i=1}^m \bar{w}_i - \beta_{\bar{\bar{T}}}^S$ elements of the set $V_1^e(\bar{T}) \cup V_2^0(\bar{T})$.

By Lemma 5.2, $|X| \geq \beta_{\bar{\bar{T}}}^S - 2m + 2$. For each $i \in X$, $w_i = 0$; therefore, it follows from Lemma 5.1 that $\bar{w}_i$ is an even number. Hence, V^i contains at least one element of $V_1^e(\bar{T}) \cup V_2^o(\bar{T})$. Thus,

$$|V_1^e(\bar{T})| + |V_2^o(\bar{T})| \geq \sum_{i=1}^m \bar{w}_i - \beta_{\bar{\bar{T}}}^S + |X| \geq \sum_{i=1}^m \bar{w}_i - \beta_{\bar{\bar{T}}}^S + \beta_{\bar{\bar{T}}}^S - 2m + 2 \geq \sum_{i=1}^m w_i - 2m + 2.$$

And $\lambda(\bar{T}) \geq \frac{|V_1| + \sum_{i=1}^m w_i - 2m + 2}{2} = \lambda(T^*)$.

In each iteration of the recursive process, all of the operations can be done in $O(n)$ time. The total number of iterations is $O(n)$. Hence, the overall computational complexity of the algorithm is $O(n^2)$. This proves the theorem. ∎

6. Subproblem 4: The optimal cut-tree realization problem (OCRP).

Statement of the problem. Given an edge-capacitated tree $T = [V, E, u]$ with integer edge capacity $u_e > 1$ for all $e \in E$, construct a network $G^* = [V, E^*, u^*]$ that has T as its cut-tree such that (i) u_e^* is integer for all $e \in E^*$ and (ii) $u^*[V] = \sum_{e \in E^*} u_e^*$ minimum.

We first establish a lower bound on the optimal objective function value of the OCRP and then we give an algorithm that achieves this bound. Thus, for any node $i \in V$, let $\pi_i = \max\{u_e : e = (i, j) \in E\}$. Let $V_1 = \{i \in V : \pi_i \text{ is odd}\}$ and $V_2 = V - V_1$. Let $E_1 = \{e : e \in E; u_e \text{ is odd}\}$ and $E_2 = E - E_1$. As before, we define $V_1^e(T)$ ($V_2^o(T)$) as the set of nodes in V_1 (V_2), each of which has an even (odd) number of odd edges (edges with capacity u_e odd) incident to it in T.

LEMMA 6.1. *Suppose a network $\bar{G} = [V, \bar{E}, \bar{u}]$, with integer-valued edge capacities $\{\bar{u}_e : e \in \bar{E}\}$, has $T = [V, E, u]$ as its cut-tree. Then*

$$\bar{u}[V] \geq \frac{\sum_{i \in V} \pi_i + |V_1^e(T)| + |V_2^o(T)|}{2}.$$

Proof. Since $\bar{G}$ has T as its cut-tree, $\bar{G}$ is flow-equivalent to T; therefore, $\delta_{\bar{u}}[i] \geq \pi_i$ for each $i \in \{1, 2, \ldots, n\}$.

Consider any node $v \in V_1^e(T) \cup V_2^o(T)$. Let the set of odd edges incident to v in T be $E_v = \{e_1, \ldots, e_\ell\}$, and let the set of even edges incident to it in T be $F_v = \{f_1, \ldots, f_q\}$. For each $e_i \in E_v$, $i = 1, \ldots, \ell$, if we delete e_i from T, we get two subtrees. Let Y_i be the node set of the subtree not containing v. Similarly, for each $f_i \in F_v$, define Z_i as the node set of the subtree of T obtained by deleting edge f_i that does not contain the node v.

In $\bar{G}$, contract node sets $Y_1, \ldots, Y_\ell$, $Z_1, \ldots, Z_q$ to nodes $y_1, \ldots, y_\ell$, $z_1, \ldots, z_q$, respectively, to get a network $\tilde{G} = [\tilde{V}, \tilde{E}, \tilde{u}]$ on node set $\tilde{V} = \{v, y_1, \ldots, y_\ell, z_1, \ldots, z_q\}$. Then, $\delta_{\tilde{u}}[v] = \delta_{\bar{u}}[v]$. Since T is a cut-tree of $\bar{G}$, $\delta_{\tilde{u}}[y_i] = \delta_{\bar{u}}[Y_i] = u_{e_i}$ for each $i \in \{1, \ldots, \ell\}$ and $\delta_{\tilde{u}}[z_i] = \delta_{\bar{u}}[Z_i] = u_{f_i}$ for each $i \in \{1, \ldots, q\}$. Thus,

$$\sum_{x \in \tilde{V}} \delta_{\tilde{u}}[x] = \sum_{i=1}^{l} u_{e_i} + \sum_{i=1}^{q} u_{f_i} + \delta_{\bar{u}}[v]. \tag{6.1}$$

In (6.1), $\sum_{i=1}^{q} u_{f_i}$ and $\sum_{x \in \tilde{V}} \delta_{\tilde{u}}[x] = 2\tilde{u}[\tilde{V}]$ are both even. Hence, if $v \in V_1^e$, then $\sum_{i=1}^{\ell} u_{e_i}$ is even, implying $\delta_{\bar{u}}[v]$ is even and if $v \in V_2^o$, then $\sum_{i=1}^{l} u_{e_i}$ is odd, implying $\delta_{\bar{u}}(v)$ is odd. In either case, $\delta_{\bar{u}}[v] \geq \pi_v + 1$. Therefore, we have

$$\bar{u}[V] = \frac{1}{2} \sum_{i \in V} \delta_{\bar{u}}[v] \geq \frac{1}{2} \left(\sum_{i \in V} \pi_i + |V_1^e| + |V_2^o| \right). \qquad \square$$

We now present an algorithm for the OCRP that produces a feasible solution to the problem with sum of edge capacities equal to the lower bound established in the above lemma and hence is an optimal solution to the problem.

Before giving a formal description of our algorithm, we briefly explain the main ideas. This will facilitate understanding of the algorithm and the proof of its validity.

Given an edge-capacitated tree $T = [V, E, u]$ with an integer edge capacity $u_e > 1$ for all $e \in E$, our algorithm first defines edge capacities $\{\tilde{u}_e : e \in E\}$ as follows:

$$\tilde{u}_e = \begin{cases} u_e & \text{if } u_e \text{ is even,} \\ u_e - 1 & \text{if } u_e \text{ is odd.} \end{cases}$$

Since $\tilde{u}_e$ is even for all $e \in E$ it follows from Theorem 2.9 that the Gomory–Hu algorithm, with $\tilde{T} = [V, E, \tilde{u}]$ as the choice of the tree in Step 0 of the algorithm, outputs a network that is an optimal solution to the corresponding instance of INSP. The network constructed by the algorithm depends on the choice of the Hamiltonian cycle $\mathcal{H}$ on node set V in Step 0 of the algorithm. We choose the cycle $\mathcal{H} = \mathcal{H}^*$ such that the solution $\hat{G} = [V, \hat{E}, \hat{u}]$ obtained has the following two properties:

(I) $\tilde{T}$ is a cut-tree of $\hat{G}$.

(II) We can add to $\hat{G}$ some edge set E_p with unit capacity per edge to get a network $G^* = [V, E^*, u^*]$ such that (i) T is a cut-tree of G^* and (ii) $\sum_{e \in E^*} u^*_e$ equals the lower bound established in Lemma 6.1.

Property (I) is ensured by choosing $\mathcal{H}^*$ such that every fundamental cut of T contains precisely two edges of $\mathcal{H}^*$.

In our approach to ensure property (II) above, the case when T is a star network plays a crucial role, and we illustrate the basic idea using the following example.

Let $T = [V, E, u]$, where $V = \{1, 2, \ldots, 10\}$, $E = \{(1, i) : i = 2, \ldots, 10\}$, and

$$u_{1i} = \begin{cases} 3 & \text{if } i = 2, \ldots, 6, \\ 2 & \text{if } i = 7, \ldots, 10. \end{cases}$$

In this case the fundamental cuts of T are $\{[\{i\}, V - \{i\}] : i = 2, \ldots, 10\}$ and every Hamiltonian cycle on node set V contains precisely two edges of each of the fundamental cuts of T. To ensure property (II) we proceed as follows: We first find an optimal solution $\mathcal{P}$ to the instance of the TPCP with input T. It is easy to see that $\mathcal{P} = \{(2-1-3), (4-1-5), (1-6)\}$ is such an optimal solution. We choose $E_p = \{(i, j) : i \text{ and } j \text{ are the endnodes of a path in } \mathcal{P}\} = \{(2, 3), (4, 5)(1, 6)\}$.

Node 1 is the unique nonleaf node of T and edge $(1, 6)$ is in E_p. Hence we choose node 6 as a special node z. If node 1 is not incident to any edge in E_p, then we choose z such that $u_{1z} = \pi_1$. Now we construct $\mathcal{H}^*$ such that the two nodes incident to every other edge in E_p lie on the two subpaths formed by deleting nodes 1 and 6 from $\mathcal{H}^*$. It is easy to verify that $\mathcal{H}^* = (1-2-4-6-5-3-1)$ is one such choice. The Gomory–Hu algorithm with choice of $\tilde{T}$ and $\mathcal{H}^*$ in Step 0 outputs network $\hat{G} = [V, \hat{E}, \hat{u}]$ with $\hat{E} = \{(1, 2), (2, 4), (4, 6), (6, 5), (5, 3), (3, 1)\}$ and $\hat{u}_{ij} = 1$ for all $(i, j) \in \hat{E}$. The tree $\tilde{T}$ can be easily seen to be a cut-tree of $\hat{G}$. In fact, it can be easily verified that for any $\phi \neq Y \subset V$,

$$\delta_{\hat{u}}[Y] = \begin{cases} 2 & \text{if node set } Y \text{ forms a subpath of } \mathcal{H}^*, \\ > 2 & \text{otherwise.} \end{cases}$$

The final network $G^* = [V, E^*, u^*]$ is obtained by adding to $\hat{G}$ the edge set E_p with unit capacity per edge. It is easy to see that

$$\delta_{u^*}[\{i\}] = \begin{cases} 3 & \text{if } i = 2, \ldots, 6, \\ 2 & \text{if } i = 7, \ldots, 10. \end{cases}$$

For every nonempty proper subset Y of V, such that $\{1, 2, \ldots, 6\} \not\subseteq Y \not\subseteq \{7, \ldots, 10\}$, and the node set Y forms a subpath of $\mathcal{H}^*$, we need $\delta_{u^*}[Y] \geq 3$; therefore, the cut $[Y, \bar{Y}]$ should contain some edge in E_p. But this follows from the choice of $\mathcal{H}^*$.

Now let us consider the general case (when T is not a star network). We first find an optimal solution $\mathcal{P}$ to the instance of TPCP with input T and define

$$E_p = \{(i, j) : i \text{ and } j \text{ are the two endnodes of a path in } \mathcal{P}\}.$$

As will be clear from the proof of the validity of our algorithm, to ensure property (II), it is sufficient for the Hamiltonian path $\mathcal{H}^*$ to satisfy the following local property.

For any nonleaf node v of T, let $T^1, \ldots, T^\ell$ be the subtrees on node sets $V^1, \ldots, V^\ell$, respectively, formed by deleting node v from T. Then each node set V^i must form a subpath p^i of $\mathcal{H}^*$, and if for each i, we contract T^i in T and p^i in $\mathcal{H}^*$ to supernode y^i, then the resultant star network $\bar{T}$ and cycle $\bar{\mathcal{H}}$ must satisfy the conditions discussed previously with respect to the set $\bar{E}_p = \{(a,b) : a \neq b$ and there exists $(i,j) \in E_p$ such that $a = i$ or (for some ℓ, $a = y^\ell$ and $i \in V^\ell$) and $b = j$ or (for some r, $b = y^r$ and $j \in V^r)\}$.

Algorithm Cut-Tree-Realization

Input. An edge-capacitated tree $T = [V, E, u]$ on node set $V = \{1, 2, \ldots, n\}$ with integer-valued edge capacity $u_e > 1$ for all $e \in E$.

Output. A network $G^* = [V, E^*, u^*]$ that is an optimal solution to the instance of the OCRP.

Step 0. Find an optimal solution $\mathcal{P}$ to the instance of the TPCP with input T, node partition (V_1, V_2), and edgepartition (E_1, E_2). Arbitrarily choose a nonleaf node $v \in T$. Initialize $S = \emptyset$, $\bar{T} = T$, $i = 0$.

Phase 1: Node ordering.

(In this phase, we construct an appropriate Hamiltonian cycle $\mathcal{H}^*$ on node set V. This cycle has two important properties: (i) every fundamental cut of T contains two edges of the cycle and (ii) certain pairs of paths in $\mathcal{P}$ cross in $\mathcal{H}^*$ (i.e., the endnodes v and z of one of the paths are in the two separate subpaths obtained by deleting the endnodes s and t of the other path from $\mathcal{H}^*$, implying that the four nodes occur in $\mathcal{H}^*$ in the order of $vsztv$).

Step 1. Let $\bar{T}^{i1}, \bar{T}^{i2}, \ldots, \bar{T}^{i\ell_i}$ be the subtrees with more than one node, each resulting from deletion of node v from tree T.

1. Out of these subtrees, add to set S those that are also subtrees of $\bar{T}$.
2. In T, contract each $\bar{T}^{ij}$ $(j = 1, \ldots, \ell_i)$ to a supernode y_{ij} to get a star network $\bar{G} = [\bar{V}, \bar{E}, \bar{u}]$ on node set $\bar{V} = \{v, x_1, x_2, \ldots, x_{k_i}, y_{i1}, y_{i2}, \ldots, y_{i\ell_i}\}$, where $x_h \in V$ $(h = 1, \ldots, k_i)$, $\bar{u}_{vx_h} = u_{vx_h}$, and $\bar{u}_{vy_{ij}} = u_{vz}$, where $(v,z) \in E$ and node z is in $\bar{T}^{ij}$.
3. For each path p in $\mathcal{P}$ and each $j \in \{1, 2, \ldots, \ell_i\}$, replace the subpath of p in $\bar{T}^{ij}$ (if any) by node y_{ij}. Let $\tilde{\mathcal{P}}$ be the collection of resultant paths of positive size.
4. If there exists a path $\tilde{p} \in \tilde{\mathcal{P}}$ with v as its endnode, then choose one such path and denote its other endnode by z. Else, let z be such that $\pi_v = \bar{u}_{vz}$.
5. Construct a cycle $\Delta\mathcal{H}^i = va_1a_2 \cdots a_{k_i+\ell_i}v$ on node set $\bar{V}$ such that (1) $a_{\lceil \frac{k_i+\ell_i}{2} \rceil} = z$ and (2) for every path $\tilde{p} \in \tilde{\mathcal{P}}$ passing through node v, the two endnodes s and t of the path are in the two subpaths formed by deleting the nodes v and z from $\Delta\mathcal{H}^i$ (i.e., the nodes v, z, s, t appear in $\Delta\mathcal{H}^i$ in the order of $vsztv$.). If $i = 0$, then set $\mathcal{H}^0 = \Delta\mathcal{H}^0$ and go to Step 3.

Step 2. Delete from $\Delta\mathcal{H}^i$ the supernode from the set $\{y_{i1}, \ldots, y_{i\ell_i}\}$ that does not correspond to any subtree of $\bar{T}$ to get a path. In $\mathcal{H}^{i-1}$ replace the supernode corresponding to $\bar{T}$ by this path to get the cycle $\mathcal{H}^i$.

Step 3. If $S \neq \emptyset$, choose an element of S, denote it by $\bar{T}$, and delete it from S. Let v be the node in $\bar{T}$ that is incident to some edge not in $\bar{T}$. Increment $i = i + 1$ and go to Step 1. Else, set $\mathcal{H}^* = \mathcal{H}^i$.

Phase 2: Design of the optimal network.
Step 4. Define

$$\tilde{u}_e = \begin{cases} u_e & \forall\ e \in E \text{ such that } u_e \text{ is an even number,} \\ u_e - 1 & \forall\ e \in E \text{ such that } u_e \text{ is an odd number.} \end{cases}$$

Let $\tilde{T} = [V, E, \tilde{u}]$. Define $\tilde{r}_{ij}$ = the maximum flow value between nodes i and j in $\tilde{T}$. Construct a network $\hat{G} = [V, \hat{E}, \hat{u}]$ using Algorithm Gomory–Hu with matrix $\tilde{R} = (\tilde{r}_{ij})_{n\times n}$ as input and with $\tilde{T}$ as the choice of the maximum weight spanning tree, and with $\mathcal{H}^*$ as the choice of the cycle on node set V in Step 0 of Algorithm Gomory–Hu.

Step 5. Let $E_p = \{(i, j) : i \text{ and } j \text{ are the endnodes of some path in } \mathcal{P}\}$. Set $E^* = \hat{E} \cup E_p$.

$$u_e^* = \begin{cases} \hat{u}_e & \forall\ e \in \hat{E} - E_p, \\ \hat{u}_e + 1 & \forall\ e \in \hat{E} \cap E_p, \\ 1 & \forall\ e \in E_p - \hat{E}. \end{cases}$$

Output the network $G^* = [V, E^*, u^*]$ and stop.

The following results will be useful in proving the validity of Algorithm Cut-Tree-Realization.

Lemma 6.2. *Let $T^0 = [V^0, E^0, u^0]$ be an edge-capacitated tree. For some edge $e = (p, q) \in E^0$, let $\bar{T}$ be the subtree containing node q formed when edge e is deleted from T^0. Let $T^1 = [V^1, E^1, u^1]$ be the tree obtained by contracting in T the subtree $\bar{T}$ to a supernode y. For each $k \in \{0, 1\}$, let $\pi_i^k = \max\{u_e^k : e = (i, j) \in E^k\}$ for all $i \in V^k$, $V_1^k = \{i \in V^k : \pi_i^k \text{ is odd}\}$, $E_1^k = \{e \in E^k : u_e^k \text{ is odd}\}$, $V_2^k = V^k - V_1^k$, and $E_2^k = E^k - E_1^k$. Let $\mathcal{P}^0$ be an optimal solution to the instance of the TPCP with input $\{T^0, (V_1^0, V_2^0), (E_1^0, E_2^0)\}$. Then the path set $\mathcal{P}^1$ containing all of the paths in $\mathcal{P}^0$ with no endnode in $\bar{T}$, together with the path in $\mathcal{P}^0$ containing the edge e (if such a path exists), with its subpath in $\bar{T}$ replaced by node y is an optimal solution to the instance of the TPCP with data $\{T^1, (V_1^1, V_2^1), (E_1^1, E_2^1)\}$.*

The proof of the above lemma follows easily from the results in section 3 and is omitted.

Lemma 6.3. *Let $T = [V, E, u]$ be the input to Algorithm* Cut-Tree-Realization. *Then, for any fundamental cut $[X, \bar{X}]$ of T, X (and hence also $\bar{X}$) is the node set of a subpath of the Hamiltonian cycle $\mathcal{H}^*$ constructed in Phase* 1 *of the algorithm.*

Proof. We shall prove the result by induction on $|V| = n$. The result is obviously correct for $n \le 3$. Suppose the result is true for all $n \le k$ for some $k \ge 3$. Let us consider the case $n = k + 1$.

For $|X| = 1$ or $n - 1$ $(= k)$, the result is obviously correct. So let us suppose that $1 < |X| < n - 1$. Let $[X, \bar{X}]$ be the fundamental cut of T corresponding to an edge $e = (x, y) \in E$ with $y \in \bar{X}$.

Suppose the nonleaf node v of T selected by Algorithm Cut-Tree-Realization in Step 0 satisfies $v \in \{x, y\}$. Without loss of generality, let us assume that $v = y$. Then in Step 1 in the first iteration, the subtree of T on node set X is replaced by a supernode y_i or z_i, and in subsequent iterations, this supernode is progressively replaced by a path on node set X in $\mathcal{H}^*$.

Suppose the node v selected by the algorithm in Step 0 is neither x nor y. Without loss of generality, let us assume that $v \in \bar{X} - \{y\}$. Then in the first iteration, for some $X \subset S \subset V$, the subtree of T on node set S is contracted in Step 1 to a supernode

y_i or z_i, and in subsequent iterations, this supernode is progressively replaced by a path on node set S in $\mathcal{H}^*$. It follows easily from Lemma 6.2 and the description of the algorithm that if in T we contract the subtree on node set $\bar{S}$ to a supernode α to get a tree $T^1 = [V^1, E^1, u^1]$ and perform the algorithm with input T^1 and choose in Step 0 the same node v and path set $\mathcal{P}^1$ as defined in Lemma 6.2, then the algorithm will produce the same path on node set S. But $|V^1| \leq k$ and $[X, \bar{X} - \bar{S} \cup \{\alpha\}]$ is a fundamental cut of T^1. Hence, it follows by induction that node set X forms a subpath of this path. This proves the lemma. □

The following is a corollary of Lemma 6.3.

COROLLARY 6.4. *Let $[X, \bar{X}]$ be the fundamental cut of $T = [V, E, u]$ corresponding to some edge $e = (i, j) \in E$. Then we have the following.*

1. *The Hamiltonian cycle $\mathcal{H}^*$ constructed in Phase 1 of Algorithm* CUT-TREE-REALIZATION *contains precisely two edges of the cut.*
2. *Every cycle $\mathcal{C}^i$ constructed by Algorithm* GOMORY–HU *invoked in Step 4 of Algorithm* CUT-TREE-REALIZATION *contains either zero or two edges of the cut.*

We shall now prove the validity of Algorithm CUT-TREE-REALIZATION.

THEOREM 6.5. *For any instance $T = (V, E, u)$ of the OCRP with $u_e > 1$ for all $e \in E$, Algorithm* CUT-TREE-REALIZATION *constructs in $O(n^2)$ time an optimal network $G^* = [V, E^*, u^*]$ for the problem with optimal value*

$$u^*[V] = \frac{\sum_{i \in V} \pi_i + |V_1^e| + |V_2^o|}{2}.$$

Proof. It follows from Theorem 2.9 that the network G^* output by the algorithm has integer edge capacities. Also, the computational complexity of the algorithm can be easily seen to be $O(n^2)$.

For the tree $\tilde{T} = [V, E, \tilde{u}]$ constructed in Step 4 of the algorithm, let $\tilde{\pi}_i = \max\{\tilde{u}_e : e = (i, j) \in E\}$ for any $i \in V$. Then

$$\tilde{\pi}_i = \begin{cases} \pi_i - 1 & \forall\ i \in V_1, \\ \pi_i & \forall\ i \in V_2. \end{cases}$$

It follows from Theorem 2.9 that the network $\hat{G} = [V, \hat{E}, \hat{u}]$ constructed in Step 4 using Algorithm GOMORY–HU satisfies

$$\hat{u}[V] = \frac{1}{2} \sum_{i \in V} \tilde{\pi}_i = \frac{1}{2} \left(\sum_{i \in V} \pi_i - |V_1| \right).$$

Also, it follows from Theorem 3.2 that $|E_p| = \frac{1}{2}(|V_1| + |V_1^e| + |V_2^o|)$. Thus

$$u^*[V] = \hat{u}[V] + |E_p| = \frac{1}{2} \left(\sum_{i \in V} \pi_i + |V_1^e| + |V_2^o| \right).$$

Thus, from Lemma 6.1, G^* (if feasible) is an optimal solution to the instance of the OCRP. Let us now prove the feasibility of G^*.

For any edge $e = (s, t) \in E$, let $[X, \bar{X}]$ be the fundamental cut in T corresponding to e with $s \in X$ and $t \in \bar{X}$. By the definition of cut-tree, to prove the feasibility of G^* it is sufficient to prove that $\delta_{u^*}[X] = u_e$ and every other cut $[Y, \bar{Y}]$ separating s and t has $\delta_{u^*}[Y] \geq u_e$.

It follows from the description of Algorithm Gomory–Hu that

$$\tilde{u}_e = \sum \left\{ \theta^i : \text{nodes } s \text{ and } t \text{ lie on the cycle } \mathcal{C}^i \right\}.$$

By Corollary 6.4, it follows that cut $[X, \bar{X}]$ contains precisely two edges of each such cycle $\mathcal{C}^i$ and no edge of any other cycle. Thus,

$$\delta_{\hat{u}}[X] = \sum \left\{ 2 \cdot \frac{1}{2}\theta^i : \text{nodes } s \text{ and } t \text{ lie on the cycle } \mathcal{C}^i \right\} = \tilde{u}_e.$$

If u_e is even, then $u_e = \tilde{u}_e$. In this case, no path in $\mathcal{P}$ covers edge e, and hence, no edge of E_p lies in cut $[X, \bar{X}]$. Thus,

$$\delta_{u^*}[X] = \delta_{\hat{u}}[X] = \tilde{u}_e = u_e.$$

If u_e is odd, then $u_e = \tilde{u}_e + 1$. In this case, precisely one path in $\mathcal{P}$ covers edge e, and hence, exactly one edge of E_p lies in cut $[X, \bar{X}]$. Thus,

$$\delta_{u^*}[X] = \delta_{\hat{u}}[X] + 1 = \tilde{u}_e + 1 = u_e.$$

Now, consider any other cut $[Y, \bar{Y}]$ with $s \in Y$ and $t \in \bar{Y}$. We will show that $\delta_{u^*}[Y] \geq u_e$. Cut $[Y, \bar{Y}]$ obviously contains at least two edges of each cycle $\mathcal{C}^i$ constructed by Algorithm Gomory–Hu that contains both nodes s and t. Thus,

$$\delta_{\hat{u}}[Y] \geq \sum \left\{ 2 \cdot \frac{1}{2}\theta^i : \text{nodes } s \text{ and } t \text{ are in } \mathcal{C}^i \right\} = \tilde{u}_e.$$

If $\tilde{u}_e = u_e$, then

$$\delta_{u^*}[Y] \geq \delta_{\hat{u}}[Y] \geq \tilde{u}_e = u_e,$$

and the result is proved. So, suppose $\tilde{u}_e = u_e - 1$, and therefore u_e is odd. For each $f \in E$, $u_f \geq 2$. Since $\tilde{u}_f$ is an even integer, it is implied that $\tilde{u}_f \geq 2$. Hence, $\theta^i \geq 2$ and is an even integer for each i. If $[Y, \bar{Y}]$ contains four or more edges of any $\mathcal{C}^i$ containing both s and t, then

$$\delta_{u^*}[Y] \geq \delta_{\hat{u}}[Y] \geq \tilde{u}_e + 2 \geq u_e.$$

So, suppose $[Y, \bar{Y}]$ contains precisely two edges of each $\mathcal{C}^i$ containing both nodes s and t, and therefore of $\mathcal{C}^0 = \mathcal{H}^*$. Then node set Y forms a subpath of $\mathcal{H}^*$. Since u_e is odd, a path of $\mathcal{P}$ covers edge e.

If s and t are the two endnodes of a path in $\mathcal{P}$, then $(s, t) \in E_p$, and therefore

$$\delta_{u^*}[Y] \geq \delta_{\hat{u}}[Y] + 1 \geq \tilde{u}_e + 1 = u_e.$$

If not, then we consider three cases.

Case 1. $\{s, t\} \subseteq V_1$, and s is the endnode of a path $p^1 \in \mathcal{P}$ that contains edge (s, t).

Let the other endnode of path p^1 be z_1, and let another path $p^2 \in \mathcal{P}$ have endnodes t and z_2, where $\{z_1, z_2\} \subseteq V - \{s, t\}$. Then, node t is not a leaf node of T, and therefore in some iteration i of Phase 1, the algorithm must choose $v = t$. Let nodes s, z_1, and z_2 belong to supernodes y_{i1}, y_{i2}, and y_{i3}, respectively. Then it follows from part (5) of Step 1 of the algorithm that in cycle $\Delta\mathcal{H}^i$, nodes t, y_{i1}, y_{i2},

and y_{i3} appear in the order $ty_{i1}y_{i3}y_{i2}t$. Hence in $\mathcal{H}^*$, nodes t, s, z_1, z_2 appear in the order tsz_2z_1t. If any of (t, z_2) and (s, z_1) lies in cut $[Y, \bar{Y}]$, then we are done. Else, $s \in Y$ and $t \in \bar{Y}$ together implies a contradiction to the fact that nodes in Y form a subpath of $\mathcal{H}^*$, and the result is proved.

Case 2. $\{s, t\} \subseteq V_1$, and neither s nor t is the endnode of any path in $\mathcal{P}$ that contains edge (s, t).

Let $\{p_1, p_2, p_3\} \subseteq \mathcal{P}$ be such that the endnodes of p_1, p_2, and p_3 are, respectively, $\{s, z_1\}$, $\{t, z_2\}$, and $\{z_3, z_4\}$, and path p_3 contains edge (s, t). Without loss of generality, let $z_3 \in X$ and $z_4 \in \bar{X}$. Then since $s \in Y$ and $t \in \bar{Y}$, we have $z_1 \in Y$ and $z_2 \in \bar{Y}$ and either $\{z_3, z_4\} \subseteq Y$ or $\{z_3, z_4\} \subseteq \bar{Y}$. For else, an edge of E_p is in cut $[Y, \bar{Y}]$, and we are done.

The above implies that neither s nor t is a leaf node of T. Hence, in some iterations i and j of Phase 1, the algorithm chooses $v = s$ and $v = t$, respectively. Without loss of generality, let $i < j$.

In iteration i, let node set $\{t, z_2, z_4\}$ belong to supernode y_{i1}, and let z_3 and z_1 belong to supernodes y_{i2} and y_{i3}, respectively. Then it follows from part (5) of Step 1 of the algorithm that nodes s, y_{i1}, y_{i2}, and y_{i3} appear in cycle $\Delta\mathcal{H}^i$ in the order $sy_{i1}y_{i3}y_{i2}s$.

In iteration j, let node set $\{s, z_1, z_3\}$ belong to supernode y_{j_1}, and let z_4 and z_2 belong to supernodes y_{j_2} and y_{j_3}, respectively. Then nodes t, y_{j1}, y_{j2}, and y_{j3} appear in cycle $\Delta\mathcal{H}^j$ in the order $ty_{j1}y_{j3}y_{j2}t$.

Hence in $\mathcal{H}^*$, nodes $\{s, t, z_1, z_2, z_3z_4\}$ appear in the order $sz_3z_1z_2z_4ts$ or $sz_3z_1tz_4z_2s$. In either case, we have a contradiction to the fact that nodes in Y form a subpath of $\mathcal{H}^*$, and the result is proved.

Case 3. $\{s, t\} \cap V_2 \neq \emptyset$.

Without loss of generality, let us assume that $t \in V_2$. Then there exists an edge $f = (t, z_1)$ in E such that u_f is even and $\pi_t = \tilde{u}_f = u_f \geq u_e + 1 \geq \tilde{u}_e + 2$.

Hence, in some iteration i, Algorithm Gomory–Hu constructs a cycle $\mathcal{C}^i$ containing nodes t and z_1 but not containing node s and assigns to this cycle a capacity $\frac{1}{2}\theta^i \geq 1$. If this cycle contains any node in Y, then

$$\delta_{u^*}[Y] \geq \delta_{\hat{u}}[Y] \geq \tilde{u}_e + \theta^i > u_e,$$

and the result is proved.

Hence, let us assume that the cycle $\mathcal{C}^i$ does not contain any node in Y. This implies that $z_1 \in \bar{Y}$, and therefore the subpath of $\mathcal{H}^*$ containing nodes t and z_1 but not containing node s is in $\bar{Y}$.

Let p_1 be the path in $\mathcal{P}$ containing the edge e. Let z_2 and z_3 be the endnodes of p_1 in Y and $\bar{Y}$, respectively. We have to consider only the case when $|\{z_2, z_3\} \cap \{s, t\}| = 0$ or 1.

Node t is not a leaf node of T. Hence, in some iteration j of Phase 1, the algorithm chooses $v = t$. If $t \neq z_2$, then the cycle $\Delta\mathcal{H}^j$ formed in the jth iteration contains node t and supernodes y_{j_1}, y_{j_2}, and y_{j_3}, containing, respectively, nodes s, z_1, and z_2, in the order $ty_{j_1}y_{j_2}y_{j_3}t$. Hence, in the cycle $\mathcal{H}^*$, nodes s, t, z_1, and z_2 occur in the order sz_1z_2ts. Therefore $z_2 \in \bar{Y}$.

If $z_3 \in Y$, then the edge $(z_3, z_2) \in E_p$ lies in the cut $[Y, \bar{Y}]$, and the result is proved.

Let us consider the case $s \neq z_3 \in \bar{Y}$. In this case, node s is the endnode of some path $p_2 \in \mathcal{P}$. Let the other endnode of p_2 be z_4. Since node s is not a leaf node of T, in some iteration i of Phase 1, the algorithm chooses $v = s$ and forms the cycle

$\Delta\mathcal{H}^i$ in which node s and the supernodes y_{i_1}, y_{i_2}, and y_{i_3}, containing, respectively, the nodes t, z_3, and z_4, occur in the order $sy_{i_1}y_{i_3}y_{i_2}s$. Hence the nodes s, t, z_3, and z_4 occur in $\mathcal{H}^*$ in the order stz_4z_3s. Since $\{t, z_3\} \subseteq \bar{Y}$, this implies that $z_4 \in \bar{Y}$ and hence the edge $(s, z_4) \in E_p$ lies in the cut $[Y, \bar{Y}]$. This proves the theorem. □

7. An algorithm for the IENSP. We will now present a combinatorial, strongly polynomial algorithm for the IENSP. The algorithm invokes as subroutines the algorithms for the subproblems discussed in the previous sections.

Algorithm EXACT-SYNTHESIS

Input. A symmetric, integer, nonnegative matrix $R = (r_{ij})_{n\times n}$.

Output. A network $G^* = [V, E^*, u^*]$, that is an optimal solution to the instance of the IENSP.

Step 0. Find a maximum weight spanning tree $T = [V, E, u]$ in G^R (the complete graph on node set V with edge weight $u_e = r_{ij}$ for all $e = (i, j)$, $i \neq j$).

1. If T does not exactly realize R, then conclude that "R is not exactly realizable" and stop.
2. Otherwise, let $\bar{E} = \{e : e \in E,\ u_e \leq 1\}$ and $F = [V, E - \bar{E}, u]$. Let $T^1, T^2, \ldots, T^k$ be the connected components of F with node sets $V^1, \ldots, V^k$, respectively.

Step 1. For each $i \in \{1, 2, \ldots, k\}$, find an optimal solution $T^{*i} = [V^i, E^i, u^i]$ to the FETP with input T^i using Algorithm TREE-FINDING.

Step 2. For each $i \in \{1, 2, \ldots, k\}$, find the optimal solution $G^{*i} = [V^i, E^{*i}, u^{*i}]$ to the OCRP with input T^{*i}.

Step 3. Construct network $G^* = [V, E^*, u^*]$ with $E^* = (\cup_{i=1}^k E^{*i}) \cup \bar{E}$ and

$$u_e^* = \begin{cases} u_e^{*i} & \text{if } e \in E^{*i} \text{ for some } i \in \{1, 2, \ldots, k\}, \\ u_e & \text{if } e \in \bar{E}. \end{cases}$$

Output the network G^* and stop.

The theorem below now follows from the results in the previous sections.

THEOREM 7.1. *Algorithm* EXACT-SYNTHESIS *produces an optimal solution to the IENSP in $O(n^2)$ time.*

Proof. Suppose an input $R = (r_{ij})_{n\times n}$ to the IENSP is exactly realizable. Let $T = [V, E, u]$ be the maximum weight spanning tree in G^R computed in Step 0 of Algorithm EXACT-SYNTHESIS.

If $u_e > 1$ for all $e \in E$, then the desired result follows from Theorems 6.5 and 5.3 and the fact that the set of all maximum weight spanning trees of G^R is precisely (i) the set of all cut-trees of all possible exact realizations of G^R and (ii) the set of all the trees flow-equivalent to T.

Suppose $u_e = 0$ for some $e \in E$. Let the corresponding fundamental cut of T be $[X, \bar{X}]$. Then R being exactly realizable implies that $r_{xy} = 0$ for any $x \in X$ and $y \in \bar{X}$. Therefore in any feasible solution G to the instance of the IENSP, node sets X and $\bar{X}$ are disconnected in G. Hence, an optimum solution to the problem can be obtained by solving the subproblems on node sets X and $\bar{X}$ with input $R_{|X}$ and $R_{|\bar{X}}$, respectively. Here $R_{|S}$ is the principal submatirx of R restricted to subindex set S for any $S \subseteq V$.

So we assume $u_e \geq 1$ for all $u \in T$ in the rest of the argument. Suppose there exist $k \geq 1$ edges in T such that $u_e = 1$. Let the subtrees obtained by deleting these k edges from T be $T^1, \ldots, T^{k+1}$ on node sets $X^1, \ldots, X^{k+1}$, respectively. Then R being exactly realizable implies that, for any two nodes $x, y \in V$, $r_{xy} > 1$ if they are in the same subtree and $r_{xy} = 1$ if they are not. Hence, in any feasible solution

G to the IENSP, (i) if we contract node sets $X^1, \ldots, X^{k+1}$ to supernodes, then we must get a tree with k edges and (ii) the subnetwork of G spanned by X^i is a feasible solution to the IENSP with input $R_{|X^i}$ for every i. Now each T^i is a maximum weight spanning tree in $G^{R_{|X^i}}$ and by Theorem 2.6; each $R_{|X^i}$ is exactly realizable. Hence, it follows from Theorems 6.5 and 5.3 that each subnetwork G^{*i} constructed by the algorithm is an optimal solution to the corresponding subproblem. Therefore, the overall optimum solution can be obtained by finding optimum subnetworks for the IENSP with $R_{|X^1}, \ldots, R_{|X^{k+1}}$ as inputs, respectively, and adding to these subnetworks the k edges on T with $e_e = 1$. This proves the theorem. □

We now show that our algorithm for the IENSP leads to an algorithm for the INSP. Thus, consider an instance of the INSP with a nonnegative, integer, and symmetric matrix R as input. Let $\pi_i = \max\{r_{ij} : j \in V - \{i\}\}$ for all $i \in V$. For convenience, we will consider only the case $\pi_i > 1$ for all i. Define a matrix $\bar{R} = (\bar{r}_{ij})_{n\times n}$ as $\bar{r}_{ij} = \min\{\pi_i, \pi_j\}$ for all $i \neq j$. Then any exact realization of $\bar{R}$ is a realization of R. It is easy to see that $T^* = [V, E, u]$ with $E = \{(1, j) : j \in V - \{1\}\}$ is an optimal solution to the corresponding FETP with $\lambda(T^*) = \lceil \frac{|V_1|}{2} \rceil$. Thus, by Theorem 7.1 Algorithm Exact-Synthesis produces an exact realization $G^* = [V, E^*, u^*]$ of $\bar{R}$ with integer capacities and $\sum_{e\in E^*} u^*_e = \lceil \frac{1}{2} \sum_{i\in V} \pi_i \rceil$. We thus get the following corollary.

Corollary 7.2. *If the input matrix R satisfies $\pi_i = \max\{r_{ij} : j \in V - \{i\}\} > 1$ for all $i \in V$, then the optimal objective function value of the corresponding instance of the INSP is $\lceil \frac{1}{2} \sum_{i\in V} \pi_i \rceil$.*

Acknowledgments. We thank the associate editor, A. Schrijver, and two anonymous referees for their valuable suggestions which greatly improved the presentation of this paper.

REFERENCES

[1] Y. P. Aneja, R. Chandrasekaran, S. N. Kabadi, and K. P. K. Nair, *Flows over edge-disjoint mixed multipaths and applications*, Discrete Appl. Math., 155 (2007), pp. 1979–2000.

[2] R. Chandrasekaran, K. P. K. Nair, Y. P. Aneja, and S. N. Kabadi, *Multi-terminal multipath flows: Synthesis*, Discrete Appl. Math., 143 (2004), pp. 182–193.

[3] W. Chou and H. Frank, *Survivable communication networks and the terminal capacity matrix*, IEEE Trans. Circuit Theory, CT-17 (1970), pp. 192–197.

[4] G. Cornuejols, D. Naddef, and W. R. Pulleyblank, *Halin graphs and the traveling salesman problem*, Math. Programming, 26 (1983), pp. 287–294.

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

[6] A. Frank, *Connectivity augmentation problems in network design*, in Mathematical Programming: State of the Art, J. Birge and K. G. Murty, eds., The University of Michigan Press, Ann Arbor, MI, 1994, pp. 34–63.

[7] R. J. Gibbens and F. P. Kelly, *Dynamic routing in fully connected networks*, IMA J. Math. Control Inform., 7 (1990), pp. 77–111.

[8] R. E. Gomory and T. C. Hu, *An application of generalized linear programming to network flows*, J. Soc. Indust. Appl. Math., 10 (1961), pp. 260–283.

[9] R. E. Gomory and T. C. Hu, *Multi-terminal network flows*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 551–570.

[10] D. Gusfield, *Simple constructions for the multi-terminal network flow synthesis*, SIAM J. Comput., 12 (1983), pp. 157–165.

[11] R. Hassin and A. Levin, *Synthesis of 2-commodity flow networks*, Math. Oper. Res., 29 (2004), pp. 280–288.

[12] M. Hojati, *The network synthesis problem in a cycle*, Oper. Res. Lett., 17 (1995), pp. 231–236.

[13] S. N. Kabadi, *Strongly Polynomial Algorithms for the Continuous and Integer Versions of the Network Augmentation Problem in a Cycle*, Technical report, Faculty of Business Administration, University of New Brunswick, New Brunswick, Canada, 2003.
[14] S. N. Kabadi, R. Chandrasekaran, and K. P. K. Nair, 2-*Commodity Integer Network Synthesis Problem*, Technical report, Faculty of Business Administration, University of New Brunswick, New Brunswick, Canada, 2003.
[15] S. N. Kabadi, R. Chandrasekaran, and K. P. K. Nair, *Multiroute flows: Cut-trees and realizability*, Discrete Optim., 2 (2005), pp. 229–240.
[16] S. N. Kabadi, R. Chandrasekaran, K. P. K. Nair, and Y. P. Aneja, *Integer version of the multipath flow network synthesis problem*, Discrete Appl. Math., to appear.
[17] S. N. Kabadi, J. Kang, R. Chandrasekaran, and K. P. K. Nair, *Hop-Constrained Network Flows: Analysis and Synthesis*, Technical report, Faculty of Business Administration, University of New Brunswick, New Brunswick, Canada, 2003.
[18] S. N. Kabadi and R. Sridhar, *Peeling Algorithm for Integral Network Synthesis*, Technical report, Faculty of Business Administration, University of New Brunswick, New Brunswick, Canada, 1996.
[19] W. Kishimoto, *A method for obtaining the maximum multiroute flows in a network*, Networks, 27 (1996), pp. 279–291.
[20] E. L. Lawler, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, 1976.
[21] W. Mayeda, *Terminal and branch capacity matrices of a communication net*, IRE Trans., CT-7 (1960), pp. 261–269.
[22] A. Punnen, private communication, 2005.
[23] A. Schrijver, *Combinatorial optimization: Polyhedra and efficiency. Vol.* B. *Matroids, trees, stable sets*, Algorithms Combin. 24, Springer-Verlag, Berlin, 2003, pp. 1049–1057.
[24] R. Sridhar and R. Chandrasekaran, *Integer solution to synthesis of communication network*, Math. Oper. Res., 17 (1992), pp. 581–585.
[25] K. Talluri, *Network synthesis with few edges*, Networks, 27 (1996), pp. 109–115.
[26] E. Tardos, *A strongly polynomial algorithm to solve combinatorial linear programs*, Oper. Res., 34 (1986), pp. 250–256.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 155–162

RAMSEY-TYPE PROBLEM FOR AN ALMOST MONOCHROMATIC K_4*

JACOB FOX† AND BENNY SUDAKOV‡

Abstract. In this short note we prove that there is a constant c such that every k-edge-coloring of the complete graph K_n with $n \geq 2^{ck}$ contains a K_4 whose edges receive at most two colors. This improves on a result of Kostochka and Mubayi, and is the first exponential bound for this problem.

Key words. Ramsey-type problems, dependent random choice, probabilistic method

AMS subject classifications. 05C55, 05C35, 05D10, 05D40

DOI. 10.1137/070706628

1. Introduction. The *Ramsey number* $R(t;k)$ is the least positive integer n such that every k-coloring of the edges of the complete graph K_n contains a monochromatic K_t. In 1916 Schur showed that $R(3;k)$ is at least exponential in k and at most a constant times $k!$. Despite various efforts over the past century to determine the asymptotics of $R(t;k)$, there were improvements only in the exponential constant in the lower bound and the constant factor in the upper bound. It is a major open problem to determine whether there is a constant c such that $R(3;k) \leq c^k$ for all k (see, e.g., the monograph [9]).

In 1981, Erdős [6] proposed studying the following generalization of the classical Ramsey problem. Let p, q be positive integers with $2 \leq q \leq \binom{p}{2}$. A (p,q)-coloring of K_n is an edge-coloring such that every copy of K_p receives at least q distinct colors. Let $f(n,p,q)$ be the minimum number of colors in a (p,q)-coloring of K_n. Determining the numbers $f(n,p,2)$ is equivalent to determining the multicolor Ramsey numbers $R(p;k)$, as an edge-coloring is a $(p,2)$-coloring if and only if it does not contain a monochromatic K_p. Over the last two decades, the study of $f(n,p,q)$ drew a lot of attention. Erdős and Gyárfás [7] proved several results on $f(n,p,q)$; e.g., they determined for which fixed p and q we have where $f(n,p,q)$ is at least linear in n, quadratic in n, or $\binom{n}{2}$ minus a constant. For fixed p, they also gave bounds on the smallest q for which $f(n,p,q)$ is asymptotically $\binom{n}{2}$. These bounds were significantly tightened by Sárközy and Selkow [15] using Szemerédi's regularity lemma. In a different paper, Sárközy and Selkow [14] show that $f(n,p,q)$ is linear in n for at most $\log p$ values of q. (Here, and throughout the paper, all logarithms are base 2.) There are also results on the behavior of $f(n,p,q)$ for particular values of p and q. Mubayi [13] gave an explicit construction of an edge-coloring which together with the already known lower bound shows that $f(n,4,4) = n^{1/2+o(1)}$. Using Behrend's construction of a dense set with no arithmetic progressions of length three, Axenovich [2] showed

*Received by the editors October 28, 2007; accepted for publication (in revised form) July 23, 2008; published electronically November 14, 2008.
http://www.siam.org/journals/sidma/23-1/70662.html

†Department of Mathematics, Princeton University, Princeton, NJ 08544 (jacobfox@math.princeton.edu). This author's research was supported by an NSF Graduate Research Fellowship and a Princeton Centennial Fellowship.

‡Department of Mathematics, UCLA, Los Angeles, CA 90095 and Institute for Advanced Study, Princeton, NJ 08544 (bsudakov@math.ucla.edu). This author's research was supported in part by NSF CAREER award DMS-0546523, NSF grants DMS-0355497 and DMS-0635607, by a USA-Israeli BSF grant, and by the State of New Jersey.

that $\frac{1+\sqrt{5}}{2}n-3 \le f(n,5,9) \le 2n^{1+c/\sqrt{\log n}}$. These examples demonstrate that special cases of $f(n,p,q)$ lead to many interesting problems.

As was pointed out by Erdős and Gyárfás [7], one of the most intriguing problems among the small cases is the behavior of $f(n,4,3)$. This problem can be rephrased in terms of another more convenient function. Let $g(k)$ be the largest positive integer n for which there is a k-edge-coloring of K_n, in which every K_4 receives at least three colors, i.e., for which $f(n,4,3) \le k$. Restated, $g(k)+1$ is the smallest positive integer n for which every k-edge-coloring of the edges of K_n contains a K_4 that receives at most two colors. In 1981, Erdős [6] showed that $g(k)$ is superlinear in k by an easy application of the probabilistic method. Later, Erdős and Gyárfás used the Lovász local lemma to show that $g(k)$ is at least quadratic in k. Mubayi [12] improved these bounds substantially, showing that $g(k) \ge 2^{c(\log k)^2}$ for some absolute positive constant c. On the other hand, the progress on the upper bound was much slower. Until very recently, the best result was of the form $g(k) < k^{ck}$ for some constant c, which follows trivially from the multicolor k-color Ramsey number for K_4. This bound was improved by Kostochka and Mubayi [10], who showed that $g(k) < (\log k)^{ck}$ for some constant c. Here we further extend their neat approach and obtain the first exponential upper bound for this problem.

THEOREM 1.1. *For $k > 2^{100}$, we have $g(k) < 2^{2000k}$.*

While it is a longstanding open problem to determine whether $R(t;k)$ grows faster than exponential in k, it is not difficult to prove an exponential upper bound if we restrict the colorings to those that do not contain a rainbow K_s for fixed s. Let $M(k,t,s)$ be the minimum n such that every k-edge-coloring of K_n has a monochromatic K_t or a rainbow K_s. Axenovich and Iverson [4] showed that $M(k,t,3) \le 2^{kt^2}$. We improve on their bound by showing that $M(k,t,s) \le s^{4kt}$ for all k,t,s. In the other direction, we prove that for all positive integers k and t with k even and $t \ge 3$, $M(k,t,3) \ge 2^{kt/4}$, thus determining $M(k,t,3)$ up to a constant factor in the exponent.

The rest of this paper is organized as follows. In the next section, we prove our main result, Theorem 1.1. In section 3, we study the Ramsey problem for colorings without rainbow K_s. The last section of this note contains some concluding remarks. Throughout the paper, we systematically omit floor and ceiling signs whenever they are not crucial for the sake of clarity. We also do not make any serious attempt to optimize absolute constants in our statements and proofs.

2. Proof of Theorem 1.1. Our proof develops further on ideas in [10]. Like the Kostochka–Mubayi proof, we show that the K_4 we find is monochromatic or is a C_4 in one color and a matching in the other color. Call a coloring of K_t *rainbow* if all $\binom{t}{2}$ edges have different colors. Let $g(k,t)$ be the largest positive integer n such that there is a k-edge-coloring of K_n with no rainbow K_t, and in which the edges of every K_4 have at least three colors. We will study $g(k)$ by investigating the behavior of $g(k,t)$.

Before jumping into the details of the proof of Theorem 1.1, we first outline the proof idea. Note that $g(k) = g(k,k)$ for $k > 2$ as a rainbow K_k would use $\binom{k}{2} > k$ colors. We give a recursive upper bound on $g(k,t)$ which implies Theorem 1.1. We first prove a couple of lemmas which show that in any k-edge-coloring without a rainbow K_t, there are many vertices that have large degree in some color i. We then apply a simple probabilistic lemma to find a large subset V_2 of vertices such that every vertex subset of size d (with $d \ll t$) has many common neighbors in color i. We use this to get an upper bound on $g(k,t)$ as follows. Consider a k-edge-coloring of K_n with $n = g(k,t)$ without a rainbow K_t and with every K_4 containing at least

three colors. There are two possible cases. If there is no rainbow K_d in the set V_2, then we obtain an upper bound on $g(k,t)$ using the fact that $|V_2|$ has size at most $g(k,d)$. If there is a set $R \subset V_2$ of d vertices which forms a rainbow K_d, then the $\binom{d}{2}$ colors that appear in this rainbow K_d cannot appear in the edges inside the set $N_i(R)$ of vertices that are adjacent to every vertex in R in color i, for otherwise we would obtain a K_4 having at most two colors (the color i and the color that appears in both R and in $N_i(R)$). In this case we obtain an upper bound on $g(k,t)$ using the fact that $|N_i(R)| \le g(k - \binom{d}{2}, t)$. Finally, if the coloring has no rainbow K_d with d constant, it is easy to show an exponential upper bound.

For an edge-coloring of K_n, a vertex x, and a color i, let $d_i(x)$ denote the degree of vertex x in color i. Our first lemma shows that if, for every vertex x and color i, $d_i(x)$ is not too large, then the coloring contains many rainbow cliques.

LEMMA 2.1. *If an edge-coloring of the complete graph K_n satisfies $d_i(x) \le \delta n$ for each $x \in V(K_n)$ and each color i, then this coloring has at most $\frac{5}{8}\delta t^4 \binom{n}{t}$ nonrainbow copies of K_t.*

Proof. If a K_t is not rainbow, then it has two adjacent edges of the same color or two nonadjacent edges of the same color. We will use this fact to give an upper bound on the number of K_t's that are not rainbow.

Let $\nu(i,t,n)$ be the number of copies of K_t in K_n in which there are at least two adjacent edges of color i. To bound the number of such K_t we can first choose the vertex, then the two edges with color i incident to this vertex and then the remaining $t-3$ vertices. Hence, the number of K_t's for which there is a vertex with degree at least two in some color is at most

$$\sum_i \nu(i,t,n) \le \sum_i \sum_{x \in V} \binom{d_i(x)}{2} \binom{n-3}{t-3} \le n\delta^{-1} \binom{\delta n}{2} \binom{n-3}{t-3}$$
$$\le \frac{\delta n^3}{2} \left(\frac{t}{n}\right)^3 \binom{n}{t} = \frac{1}{2}\delta t^3 \binom{n}{t}.$$

Here we used the fact that $\sum_i \binom{d_i(x)}{2} \le \delta^{-1}\binom{\delta n}{2}$, since $d_i(x) \le \delta n$, $\sum_i d_i(x) = n-1$, and the function $f(y) = \binom{y}{2}$ is convex.

Let $\psi(i,t,n)$ be the number of copies of K_t in K_n in which there is a matching of size at least two in color i. Let e_i denote the number of edges of color i. Since

$$e_i \le \frac{n}{2} \max_{x \in V} d_i(x) \le \frac{\delta}{2} n^2,$$

then the number of K_t's in which there is a matching of size at least two in some color is at most

$$\sum_i \psi(i,t,n) \le \sum_i \binom{e_i}{2} \binom{n-4}{t-4} \le \delta^{-1} \binom{\delta n^2/2}{2} \binom{n-4}{t-4} \le \frac{\delta t^4}{8} \binom{n}{t},$$

where again we used the convexity of the function $f(y) = \binom{y}{2}$ together with $e_i \le \delta n^2/2$ and $\sum_i e_i \le n^2/2$. Hence, the number of K_t's which are not rainbow is at most $\frac{1}{2}\delta t^3 \binom{n}{t} + \frac{1}{8}\delta t^4 \binom{n}{t} \le \frac{5}{8}\delta t^4 \binom{n}{t}$, completing the proof. ∎

For the proof of Theorem 1.1, we do not need the full strength of this lemma since we will use only the existence of at least one rainbow K_t. We also would like to mention the following stronger result. Call an edge-coloring *m-good* if each color appears at most m times at each vertex. Let $h(m,t)$ denote the minimum n such

that every m-good edge-coloring of K_n contains a rainbow K_t. The above lemma demonstrates that $h(m,t)$ is at most mt^4. It is shown by Alon et al. [1] that there are constant positive constants c_1 and c_2 such that

$$c_1 mt^3/\log t \le h(m,t) \le c_2 mt^3/\log t.$$

The following easy corollary of Lemma 2.1 demonstrates that in every k-edge-coloring without a rainbow K_t, there is a color and a large set of vertices which have large degree in that color.

COROLLARY 2.2. *In every k-edge-coloring of K_n without a rainbow K_t, there is a subset $V_1 \subset V(K_n)$ with $|V_1| \ge \frac{n}{2k}$ and a color i such that $d_i(x) \ge \frac{n}{2t^4}$ for each vertex $x \in V_1$.*

Proof. Let $V' \subset V(K_n)$ be those vertices x for which there is a color i such that $d_i(x) \ge \frac{n}{2t^4}$.

Case 1: $|V'| < n/2$. In this case, letting $V'' = V(K_n) \setminus V'$, $|V''| \ge n/2$ and no vertex in V'' has degree at least $\frac{n}{2t^4} \le |V''|/t^4$ in any given color. By Lemma 2.1 applied to the coloring of K_n restricted to V'' with $\delta = t^{-4}$, there are at least $\frac{3}{8}\binom{|V''|}{t}$ rainbow K_t's, contradicting the assumption that the coloring is free of rainbow K_t's.

Case 2: $|V'| \ge n/2$. In this case, by the pigeonhole principle, there is a color i and at least $\frac{n}{2k}$ vertices x for which $d_i(x) \ge \frac{n}{2t^4}$, completing the proof. ∎

The following lemma is essentially the same as results in [11] and [16]. Its proof uses a probabilistic argument commonly referred to as dependent random choice, which appears to be a powerful tool in proving various results in Ramsey theory (see, e.g., [8] and its references). In a graph G, the *neighborhood* $N(v)$ of a vertex v is the set of vertices adjacent to v. For a vertex subset U of a graph G, the *common neighborhood* $N(U)$ is the set of vertices adjacent to all vertices in U.

LEMMA 2.3. *Let $G = (V,E)$ be a graph with n vertices and let $V_1 \subset V$ be a subset with $|V_1| = m$ in which each vertex has degree at least αn. If $\beta \le m^{-d/h}$, then there is a subset $V_2 \subset V_1$ with $|V_2| \ge \alpha^h m - 1$ such that every d-tuple in V_2 has at least βn common neighbors.*

Proof. Let $U = \{x_1, \dots, x_h\}$ be a subset of h random vertices from V chosen uniformly with repetitions, and let $V_1' = N(U) \cap V_1$. We have

$$\mathbb{E}[|V_1'|] = \sum_{v \in V_1} \mathbf{Pr}(v \in N(U)) = \sum_{v \in V_1} \left(\frac{|N(v)|}{n}\right)^h \ge \alpha^h m.$$

The probability that a given set $W \subset V_1$ of vertices is contained in V_1' is $(\frac{|N(W)|}{n})^h$. Let Z denote the number of d-tuples in V_1' with less than βn common neighbors. So

$$\mathbb{E}[Z] = \sum_{W \subset V_1, |W|=d, |N(W)|<\beta n} \mathbf{Pr}(W \subset V_1') \le \binom{m}{d}\beta^h \le m^d \beta^h \le 1.$$

Hence, the expectation of $|V_1'| - Z$ is at least $\alpha^h m - 1$ and thus, there is a choice U_0 for U such that the corresponding value of $|V_1'| - Z$ is at least $\alpha^h m - 1$. For every d-tuple D of vertices of V_1' with less than βn common neighbors, delete a vertex $v_D \in D$ from V_1'. Letting V_2 be the resulting set, it is clear that V_2 has the desired properties, completing the proof. ∎

The proof of the next lemma uses the standard pigeonhole argument together with Lemma 2.1.

LEMMA 2.4. *Let d, k be integers with $d, k \geq 2$. Then every k-edge-coloring of K_n with $n \geq d^{12k}$ and without a rainbow K_d has a monochromatic K_4. In particular, we have $g(k,d) < d^{12k}$.*

Proof. Suppose for contradiction that there is a k-edge-coloring of K_n with $n \geq d^{12k}$ and without a rainbow K_d and without a monochromatic K_4. By Lemma 2.1 with $t = d$ and $\delta = d^{-4}$, this graph contains a vertex x_1 with degree at least $\frac{n}{d^4}$ in some color c_1. Pick this vertex x_1 and let N_1 be the set of vertices adjacent to x_1 by color c_1. We will define a sequence $x_1, \dots, x_{2k+1}$ of vertices, a sequence $c_1, \dots, c_{2k+1}$ of colors, and a sequence $V(K_n) \supset N_1 \supset \cdots \supset N_{2k+1}$ of vertex subsets. Once x_j, c_j, and N_j have been defined, pick a vertex x_{j+1} in N_j such that there are at least $\frac{|N_j|}{d^4}$ vertices in N_j connected to x_{j+1} by edges of the same color c_{j+1}. Let N_{j+1} be the set of vertices in N_j that are adjacent to x_{j+1} by edges of color c_{j+1}. Note that $|N_{j+1}| \geq d^{-4}|N_j|$ so

$$|N_{2k+1}| \geq (d^{-4})^{2k+1} n \geq 1.$$

Therefore, there is a color c that is represented at least three times in the list $c_1, \dots, c_{2k+1}$ and the three vertices $x_{j_1}, x_{j_2}, x_{j_3}$ together with a vertex from N_{2k+1} form a monochromatic K_4 in color c, where $c_{j_1} = c_{j_1} = c_{j_3} = c$ with $j_1 < j_2 < j_3$. ∎

LEMMA 2.5. *Let d, k, t be positive integers with $3 \leq d \leq t$ and $d \geq 40 \log t$. If $k \geq \binom{d}{2}$, then*

$$g(k,t) \leq \max\left(4kg(k,t)^{\frac{20\log t}{d}} g(k,d), 2^{\binom{d}{2}} g\left(k - \binom{d}{2}, t\right)\right). \tag{2.1}$$

Otherwise, we have $g(k,t) = g(k,d)$.

Proof. Note that if $k < \binom{d}{2}$, then a k-edge-coloring cannot have a rainbow K_d. Therefore, $g(k,t) = g(k,d)$ in this case. So we assume $k \geq \binom{d}{2}$. By the definition of $g(k,t)$, there is a k-edge-coloring of K_n with $n = g(k,t)$ with no rainbow K_t and in which every K_4 receives at least three colors. Consider such a coloring. By Corollary 2.2, there is a color i and a subset $V_1 \subset V(K_n)$ with $|V_1| \geq \frac{n}{2k}$ and $d_i(x) \geq \frac{n}{2t^4}$ for every vertex $x \in V_1$. Apply Lemma 2.3 to the graph of color i with $\alpha = \frac{1}{2t^4}$, $\beta = 2^{-\binom{d}{2}}$, $m = |V_1| \geq \frac{n}{2k}$, and $h = 4d^{-1}\log n$. We can apply Lemma 2.3 since $\beta < 2^{-d^2/4} = n^{-d/h} \leq |V_1|^{-d/h}$. So there is a subset $V_2 \subset V_1$ such that

$$|V_2| \geq \alpha^h m - 1 \geq \alpha^h m/2 \geq (2t^4)^{-4d^{-1}\log n} \cdot \frac{n}{4k} \geq n^{1-\frac{20\log t}{d}}/(4k)$$

and every subset of V_2 of size d has at least $\beta n = 2^{-\binom{d}{2}} n$ common neighbors in color i.

There are two possibilities: Either every K_d in V_2 is not rainbow, or there is a K_d in V_2 that is rainbow. In the first case, the k-edge-coloring restricted to V_2 is free of rainbow K_d, so

$$g(k,d) \geq |V_2| \geq n^{1-\frac{20\log t}{d}}/(4k).$$

Since $n = g(k,t)$, we can restate this inequality as

$$g(k,t) \leq 4kg(k,t)^{\frac{20\log t}{d}} g(k,d).$$

In the second case, there is a rainbow d-tuple $R \subset V_2$ such that $N_i(R)$, the common neighborhood of R in color i, has cardinality at least βn. The $\binom{d}{2}$ colors present in R

cannot be present in $N_i(R)$ since otherwise we would have a K_4 using only two colors (the color i and the color that appears in both R and in $N_i(R)$). In this case we have

$$g\left(k-\binom{d}{2},t\right)\geq|N_i(R)|\geq\beta n=2^{-\binom{d}{2}}g(k,t).$$

In either case we have

$$g(k,t)\leq\max\left(4kg(k,t)^{\frac{20\log t}{d}}g(k,d),2^{\binom{d}{2}}g\left(k-\binom{d}{2},t\right)\right),$$

which completes the proof. □

Having finished all the necessary preparation, we are now ready to prove Theorem 1.1, which says that $g(k)\leq 2^{2000k}$ for $k>2^{100}$. The iterated logarithm $\log^* n$ is defined by $\log^* n=0$ if $n\leq 1$ and otherwise $\log^* n=1+\log^*\log n$. It is straightforward to verify that $\log^* n<\log n$ holds for $n>8$.

Proof of Theorem 1.1. Note that $g(k)=g(k,k)$ since no k-edge-coloring contains a rainbow K_k. Assume $k>2^{100}$ and suppose for contradiction that there is a k-edge-coloring of K_n with $n=g(k)\geq 2^{2000k}$ such that every K_4 has at least three colors.

Let $t_1=k$, and if $t_i>2^{100}$, let $t_{i+1}=(\log t_i)^2$. We first exhibit several inequalities which we will use. We have $t_{i+1}>100\log t_i$ and $20\frac{\log t_i}{t_{i+1}}=20/\log t_i\leq\frac{1}{5}$. Let ℓ be the largest positive integer for which t_ℓ is defined, so $100^2<t_\ell\leq 2^{100}$. Note that $\ell<2\log^* k$, as one can easily check that $t_{j+1}=(\log t_j)^2=(2\log\log t_{j-1})^2<\log t_{j-1}$. Since $\ell<2\log^* k\leq 2\log k$ and $n\geq 2^{2000k}$, then $(4k)^\ell<n^{1/12}$. For $1\leq i\leq\ell-1$, we have $20/\log t_{\ell-i}<5^{-i}$. Indeed, for $i=1$, since $t_{\ell-1}>2^{100}$, we have $20/\log t_{\ell-1}<1/5$. Suppose by induction on i that we already have $20/\log t_{\ell-i}<5^{-i}$. Then $t_{\ell-i}>2^{20\cdot 5^i}$ and therefore we have $20/\log t_{\ell-i-1}=20/\sqrt{t_{\ell-i}}\leq 20\cdot 2^{-10\cdot 5^i}<5^{-i-1}$. Therefore, $\sum_{i=1}^{\ell-1}20/\log t_i<\sum_{i=1}^{\infty}5^{-i}\leq 1/4$. Putting this together, we have

$$(4k)^{\ell-1}n^{\sum_{i=1}^{\ell-1}20/\log t_i}<n^{1/3}.$$

To get an upper bound on $g(k,k)$ we repeatedly apply Lemma 2.5. Given $k'\leq k$ and $t=t_i$, to bound $g(k',t)$, we use this lemma with $d=t_{i+1}$. Note that we have $d=t_{i+1}>100\log t_i$, so indeed the condition of the lemma holds. If $k'<\binom{t_{i+1}}{2}$, then $g(k',t_i)=g(k',t_{i+1})$. Otherwise, we have one of two possible upper bounds given by (2.1). If the maximum of the two terms in (2.1) is the left bound, then

$$g(k',t)\leq 4k'g(k',t)^{\frac{20\log t}{d}}g(k',d)\leq 4kn^{\frac{20\log t}{d}}g(k',d)=4kn^{20/\log t_i}g(k',d);$$

otherwise we have $g(k',t)\leq 2^j g(k'-j,t)$ with $j=\binom{d}{2}$. Since $\frac{g(k',t)}{g(k',d)}\leq 4kn^{20/\log t_i}$ if the left bound holds, we can accumulate only up to a total upper bound factor of

$$\prod_{i=1}^{l-1}4kn^{20/\log t_i}=(4k)^{\ell-1}n^{\sum_{i=1}^{\ell-1}20/\log t_i}<n^{1/3}$$

in all of the applications of the left bound. When we use the right bound, we pick up a factor of $\frac{g(k',t)}{g(k'-j,t)}\leq 2^j$ with $j=\binom{d}{2}$ and also decrease k' by j. Since in the end of the process $k'\geq 3$, this can give only another multiplicative factor of at most 2^k in all of the applications of the right bound.

As we already mentioned above, if $k' < \binom{t_{i+1}}{2}$, then $g(k', t_i) = g(k', t_{i+1})$. Therefore when we finish repeatedly applying Lemma 2.5 we end up with a term of the form $g(k_0, t_\ell)$ with $k_0 \le k$. In that case, we use that $t_\ell \le 2^{100}$ together with Lemma 2.4 to bound it by $g(k, t_\ell) \le t_\ell^{12k} \le 2^{1200k}$. Putting this all together, we obtain the upper bound

$$n = g(k) = g(k,k) < n^{1/3} 2^k g(k, t_\ell) < 2^{1201k} n^{1/3},$$

which implies that $n < 2^{2000k}$. This completes the proof. ∎

3. Monochromatic or rainbow cliques. In this section, we prove bounds on the smallest n, denoted by $M(k,t,s)$, such that every k-edge-coloring of K_n contains a monochromatic K_t or a rainbow K_s. The following proposition is a straightforward generalization of Lemma 2.4.

PROPOSITION 3.1. *We have* $M(k,t,s) \le s^{4kt}$.

Let $M_s(t_1, \dots, t_k)$ be the maximum n such that there is a k-edge-coloring of K_n with colors $\{1, \dots, k\}$ without a rainbow K_s and without a monochromatic K_{t_i} in color i for $1 \le i \le k$. The above proposition follows from repeated application of the following recursive bound.

LEMMA 3.2. *We have*

$$M_s(t_1, \dots, t_k) \le s^4 \max_{1 \le i \le k} M_s(t_1, \dots, t_i - 1, \dots, t_k).$$

Proof. By Lemma 2.1, for every edge-coloring of K_n without a rainbow K_s, there is a vertex v with degree at least n/s^4 in some color i. If the coloring of K_n does not contain a monochromatic K_{t_i} in color i, then the neighborhood of v in color i has at least n/s^4 vertices and does not contain $K_{t_i - 1}$ in color i, completing the proof. ∎

Using a slightly better estimate by Alon et al. [1] (which we mentioned earlier) instead of Lemma 2.1, one can improve the constant in the exponent of the above proposition from 4 to 3. Together with the next lemma, Proposition 3.1 determines $M(k,t,3)$ up to a constant factor in the exponent.

LEMMA 3.3. *For all positive integers k and t with k even and $t \ge 3$, we have* $M(k,t,3) > 2^{kt/4}$.

Proof. To prove the lemma, it suffices by induction to prove $M(k,t,3) - 1 \ge 2^{t/2}\,(M(k-2,t,3) - 1)$ for all $k \ge 2$ and $t \ge 3$. Consider a 2-edge-coloring C_1 of K_m with $m = 2^{t/2}$ and without a monochromatic K_t. Such a 2-edge-coloring exists by the well-known lower bound of Erdős [5] on the 2-color Ramsey number $R(t;2)$. Consider also a $(k-2)$-edge-coloring C_2 of K_r with $r = M(k-2,t,3) - 1$ without a rainbow triangle and without a monochromatic K_t. We use these two colorings to make a new edge-coloring C_3 of K_{mr} with k colors: We first partition the vertices of K_{mr} into m vertex subsets $V_1, \dots, V_m$ each of size r, and color any edge $e = (v,w)$ with $v \in V_i, w \in V_j$, and $i \ne j$ by the color of (i,j) in the 2-edge-coloring C_1 of K_m, and color within each V_i identical to the coloring C_2 of K_r. First we show that coloring C_3 has no rainbow triangle. Indeed, consider three vertices of K_{mr}. If all three vertices lie in the same vertex subset V_i, then the triangle between them is not rainbow by the assumption on coloring C_2. If exactly two of the three vertices lie in the same vertex subset, then the two edges from these vertices to the third vertex will receive the same color. Finally, if they lie in three different vertex subsets, then the triangle between them receives only colors from C_1 and is not rainbow since C_1 is a 2-coloring. Similarly, one can see that coloring C_3 has no monochromatic K_t, which completes the proof. ∎

4. Concluding remarks. In this paper we proved that there exists a constant c such that every k-edge-coloring of K_n with $n \geq 2^{ck}$ contains a K_4 whose edges receive at most two colors. On the other hand, for $n \leq 2^{c(\log k)^2}$, Mubayi constructed a k-edge-coloring of K_n in which every K_4 receives at least three colors. There is still a large gap between these results. We believe that the lower bound is closer to the truth, and the correct growth is likely to be subexponential in k.

Our upper bound is equivalent to $f(n,4,3) \geq (\log n)/2000$ for n sufficiently large. Kostochka and Mubayi showed that $f(n,2a,a+1) \geq c_a \frac{\log n}{\log\log\log n}$, where c_a is a positive constant for each integer $a \geq 2$. Like the Kostochka–Mubayi proof, our proof can be generalized to demonstrate that for every integer $a \geq 2$ there is $c_a > 0$ such that $f(n,2a,a+1) \geq c_a \log n$ for every positive integer n. For brevity, we do not include the details.

We do not yet have a good understanding of how $M(k,t,s)$, which is the smallest positive integer n such that every k-edge-coloring of K_n has a monochromatic K_t or a rainbow K_s, depends on s. From the definition, it is an increasing function in s. For constant s, we showed that $M(k,t,s)$ grows only exponentially in k. On the other hand, for $\binom{s}{2} > k$, we have $M(k,t,s) = R(t;k)$, so understanding the behavior of $M(k,t,s)$ for large s is equivalent to understanding the classical Ramsey numbers $R(t;k)$.

REFERENCES

[1] N. Alon, T. Jiang, Z. Miller, and D. Pritikin, *Properly colored subgraphs and rainbow subgraphs in edge-colorings with local constraints*, Random Structures Algorithms, 23 (2003), pp. 409–433.

[2] M. Axenovich, *A generalized Ramsey problem*, Discrete Math., 222 (2000), pp. 247–249.

[3] M. Axenovich, Z. Füredi, and D. Mubayi, *On generalized Ramsey theory: The bipartite case*, J. Combin. Theory Ser. B, 79 (2000), pp. 66–86.

[4] M. Axenovich and P. Iverson, *Edge-colorings avoiding rainbow and monochromatic subgraphs*, Discrete Math., 308 (2008), pp. 4710–4723.

[5] P. Erdős, *Some remarks on the theory of graphs*, Bull. Amer. Math. Soc., 53 (1947), pp. 292–294.

[6] P. Erdős, *Solved and unsolved problems in combinatorics and combinatorial number theory*, Congr. Numer., 32 (1981), pp. 49–62.

[7] P. Erdős and A. Gyárfás, *A variant of the classical Ramsey problem*, Combinatorica, 17 (1997), pp. 459–467.

[8] J. Fox and B. Sudakov, *Density theorems for bipartite graphs and related Ramsey-type results*, Combinatorica, to appear.

[9] R. Graham, B. Rothschild, and J. Spencer, *Ramsey Theory*, 2nd ed., Wiley, New York, 1990.

[10] A. Kostochka and D. Mubayi, *When is an almost monochromatic K_4 guaranteed?*, submitted.

[11] A. Kostochka and V. Rödl, *On graphs with small Ramsey numbers*, J. Graph Theory, 37 (2001), pp. 198–204.

[12] D. Mubayi, *Edge-coloring cliques with three colors on all 4-cliques*, Combinatorica, 18 (1998), pp. 293–296.

[13] D. Mubayi, *An explicit construction for a Ramsey problem*, Combinatorica, 24 (2004), pp. 313–324.

[14] G. N. Sárközy and S. M. Selkow, *On edge colorings with at least q colors in every subset of p vertices*, Electron. J. Combin., 8 (2001), Research Paper 9, 6 pp.

[15] G. N. Sárközy and S. M. Selkow, *An application of the regularity lemma in generalized Ramsey theory*, J. Graph Theory, 44 (2003), pp. 39–49.

[16] B. Sudakov, *Few remarks on the Ramsey-Turan-type problems*, J. Combin. Theory Ser. B, 88 (2003), pp. 99–106.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 163–177

APPROXIMATE INTEGER DECOMPOSITIONS FOR UNDIRECTED NETWORK DESIGN PROBLEMS*

CHANDRA CHEKURI† AND F. BRUCE SHEPHERD‡

Abstract. A well-known theorem of Nash-Williams and Tutte gives a necessary and sufficient condition for the existence of k edge-disjoint spanning trees in an undirected graph. A corollary of this theorem is that every $2k$–edge-connected graph has k edge-disjoint spanning trees. We show that the splitting-off theorem of Mader in undirected graphs implies a generalization of this to finding k edge-disjoint Steiner forests in Eulerian graphs. This leads to new 2-approximation rounding algorithms for certain constrained 0-1 forest problems considered by Goemans and Williamson. These algorithms also produce approximate integer decompositions of fractional solutions. We then discuss open problems and outlets for this approach to the more general class of 0-1 skew supermodular network design problems.

Key words. network design, supermodular function, integer decomposition, approximation algorithm

AMS subject classifications. 68Q25, 68W25, 90C27, 90C59

DOI. 10.1137/040617339

1. Introduction. In this article we consider the application of splitting-off techniques to obtain integer decomposition theorems and rounding algorithms for undirected network design problems such as the Steiner forest problem and others. A well-known theorem in graph theory is the following.

THEOREM 1.1 (Nash-Williams and Tutte). *Given an undirected multigraph $G = (V, E)$, there exist k edge-disjoint spanning trees $T_1, T_2, \ldots, T_k$ in G if and only if for every partition $V_1, V_2, \ldots, V_\ell$ of V the number of edges between the node sets of the partition is at least $k(\ell - 1)$.*

An easy corollary of the above is the following.

COROLLARY 1.2. *If G is $2k$–edge-connected, then there exist k edge-disjoint spanning trees in G.*

Let $\lambda_G(u, v)$ denote the connectivity between u and v in G. We consider packing Steiner forests instead of spanning trees and obtain the following generalization of Corollary 1.2 for Eulerian graphs.

LEMMA 1.3 (the forest packing lemma). *Given a Eulerian graph G and pairs of nodes $s_1t_1, s_2t_2, \ldots, s_\ell t_\ell$ such that for $1 \le i \le \ell$, $\lambda_G(s_i, t_i) \ge 2k$, there are k edge-disjoint forests $F_1, F_2, \ldots, F_k$ such that in each F_j, s_i and t_i are connected for $1 \le i \le \ell$.*

We give a proof of this result (in section 3) that relies on a simple application of Theorem 1.1 and the classical splitting-off technique of Mader [32]. A special case is proved in [17] where the goal is to pack Steiner "S-trees," i.e., trees that each contains a given subset S of the nodes. Splitting-off for packing Steiner trees in general graphs is also considered in [26]. While simple, the extension above to forests, rather than

*Received by the editors October 20, 2004; accepted for publication (in revised form) June 11, 2008; published electronically December 17, 2008. This work was done while both authors were at Lucent Bell Labs, and their work was partially funded by a basic research grant N00014-03-M-0141 from ONR to Lucent Bell Labs.

http://www.siam.org/journals/sidma/23-1/61733.html

†Dept. of Computer Science, University of Illinois, Urbana, IL 61801 (chekuri@cs.uiuc.edu).

‡McGill University, 805 Sherbrooke West, Montreal H3A 2K6, QC, Canada (bruce.shepherd@mcgill.ca).

trees, is of interest in its own right and has already been of use in a related context [29].

Our algorithmic motivation for proving Lemma 1.3 actually arises from the following network design problem, which is "dual" to the forest packing problem. In the *Steiner forest problem* (also called the generalized Steiner problem) we are given an edge-weighted undirected graph $G = (V, E, w)$ and a set of pairs $s_1t_1, s_2t_2, \ldots, s_\ell t_\ell$. The goal is to find a minimum cost subgraph H of G such that, for $1 \leq i \leq \ell$, s_i and t_i are connected in H. This problem has been studied intensively, with some of the most general outcomes appearing in [21, 25]. Ultimately we seek results for packing more general classes of subgraphs, not just forests, in connection with network design arising from certain supermodular set functions. We outline this more general framework in the following subsections and state Theorem 1.5, a strict generalization of the forest packing lemma, that we prove in this paper.

1.1. Approximate integer decomposition properties. In this article we work exclusively with the standard cut-based linear programming (LP) *relaxation* for our network design problems. For $e \in E$ there is a variable $x_e \in [0, 1]$ that indicates if e is part of the subgraph. We seek to minimize $\sum_e w_e x_e$ subject to the constraint that for each $S \subset V$ that separates some pair s_it_i, $x(\delta(S)) \geq 1$. The primal-dual 2-approximation algorithms of Agrawal, Klein, and Ravi [3] and later Goemans and Williamson [20] show that the integrality gap of the cut-based LP is $(2 - 2/h)$, where h is the number of distinct terminals. We obtain an alternative proof of a gap of 2, and of more interest, we show the relaxation (and our 2-approximation algorithm) has a stronger "integrality" property. We can describe this now.

Let x be a solution to the LP, and let k be an integer such that kx is integral. Consider the graph $G' = (V, E')$ obtained by taking $2kx_e$ copies of each edge $e \in E$. By Lemma 1.3 it follows that E' contains k edge-disjoint forests, each of which is a feasible solution to the Steiner forest problem. Thus the vector $2kx$ dominates a sum of k integral solutions. By convexity it follows that one of the k forests is of cost no more than $2w \cdot x$, in other words twice the cost of the original LP solution.

The above approach yields a 2-approximation algorithm with a stronger property than those from earlier methods in the following sense: it is always the case that if the integrality gap of an LP relaxation for a minimization problem is $\alpha \geq 1$, then αx dominates a convex (i.e., fractional) combination of integral solutions. However, in general, it does not follow that αkx, when kx is integral, dominates a sum (i.e., integral combination) of k integral solutions. If this stronger property holds for any feasible fractional solution x, we say the relaxation has the *α-approximate integer decomposition property*; more precisely, this is a property of the polyhedron consisting of feasible solutions for the relaxation. If $\alpha = 1$, the above decomposition property is called the *integer decomposition property* (IDP) and is well-studied, cf. [34]. Baum and Trotter [5, 6] show, for instance, that a matrix A is totally unimodular if and only if $\{x : Ax \leq b, x \geq 0\}$ has the IDP for each integral b. Approximate integer decomposition for maximization problems, in particular packing problems, can be defined in a fashion similar to that for minimization problems. For a fractional solution x, one considers an integral vector kx but seeks a decomposition (or cover) of kx into at most $\alpha k \geq k$ integer feasible solutions: $kx = \sum_{i=1}^{\lfloor \alpha k \rfloor} g_i$. Obviously, one of the g_i's is an integral solution whose weight (profit) is at least $\frac{1}{\alpha}$ times that of x.

This decomposition approach is perfectly natural and is often the technique used in the literature to establish an approximation ratio (the first mention of a connection to the IDP seems to appear in [10]). Some well-known combinatorial problems have

an integrality gap equal to their approximation ratio for integer decomposition. For instance, it is an exercise to show that the natural LP relaxation for the knapsack problem has the 2-approximate IDP. It is not always obvious, however, when such a property does hold. In this paper, we ask, for example, whether recently celebrated 2-approximation results of Jain [25] can be extended to have the 2-approximate IDP (see section 1.2.1).

We believe it is not only worthwhile to make the integer decomposition approach explicit (including its connections to traditional polyhedral results for IDP) but also that such stronger decomposition results are potentially important in their own right. For instance, the results of [10] for packing paths in trees provided the stronger IDP. These results were subsequently used in [1, 2, 13], where the integer decompositions corresponded to partitioning pairwise demands so that each class of demands could be routed with a distinct wavelength in an optical network. In some recent work, Fukunaga and Nagamochi [14] applied the approximate integer decomposition methodology to obtain algorithms for the set connector problem.

Before continuing with our main focus, approximations for network design, we give another application of integer decompositions, this time to yield an approximation result due to Goemans and Williamson [20] for the prize-collecting Steiner tree problem. Namely, we mention that their result can be alternatively derived from a result of Bang-Jensen, Frank, and Jackson [4] on packing arc-disjoint Steiner arborescences in directed graphs. We give some details below. In the *prize-collecting Steiner tree problem* we are given an undirected edge-weighted graph $G = (V, E, c)$ and a root node $r \in V$. Each node v also has a nonnegative penalty value $\pi(v)$. The objective is to find a tree $T = (V(T), E(T))$ rooted at r that minimizes $\sum_{e \in E(T)} c(e) + \sum_{v \notin V(T)} \pi(v)$. The first constant factor approximation algorithm for this problem was given in [7] and subsequently [20], which gave a primal-dual algorithm that finds a tree T such that $\sum_{e \in E(T)} c(e) + 2 \sum_{v \notin V(T)} \pi(v) \leq 2\text{OPT}$, where OPT is the optimum value of the natural LP relaxation for the problem. This result has found use in other approximation algorithms, notably for the k-minimum spanning tree problem [8, 18] and several others. The result can be obtained from [4] as follows: consider a fractional solution x to the LP relaxation: x_e is the value on edge e and $x(v)$ is the flow from r to v supported by x. We obtain a directed graph by bidirecting each edge e and placing a value of x_e on both of the resulting arcs. This clearly increases the cost of edges by a factor of 2. Now we apply Theorem 2.1 in [4] to obtain a convex combination of arborescences rooted at r in which each v occurs in at least $x(v)$ arborescences. Picking the lowest cost arborescence yields the desired result. The remaining details are left to the interested reader.

1.2. Constrained forest problems and f-connected networks. Goemans and Williamson [20] obtain 2-approximation algorithms for a large class of network design problems that they refer to as *constrained forest problems*; they apply their primal-dual framework for this. Each of these problems is determined by an integer-valued function f that for each set $S \subseteq V$ gives a requirement value $f(S)$. (In some cases, we only require this for sets in a given family $\mathcal{F}$—see section 1.2.1.) A solution to the connectivity problem modeled by f is a collection of edges A such that at least $|A \cap \delta(S)| \geq f(S)$ for each $S \subseteq V$. Such a solution will be called *f-connected*, or an *f-connector*. The optimization problem is to find a minimum cost f-connector.

The most general class of functions for which the network design problem is known to have a constant factor approximation is the set of integer-valued skew supermodular functions. In establishing this result, Jain [25] introduced a new iterative rounding

approach to obtain a 2-approximation for such skew supermodular problems, called *Steiner network design problems*. As we see in section 1.2.1, many natural (NP-hard or otherwise) network design problems are modeled as minimum cost f-connector problems for a skew supermodular function f. In section 5 we discuss a kind of inverse problem which we believe deserves further investigation. Given a requirement function, does it encode a natural class of network design problems? We give several results on when $\{0,1\}$-valued requirement functions encode certain connectivity augmentation design problems.

Jain's approach, based on the framework designed for submodular flows [12, 31], requires finding a *basic* solution to the cut LP relaxation for f-connected subgraphs. One of our motivations for studying primal rounding methods via a decomposition-based approach is to find a combinatorial rounding algorithm for the Steiner network problem. The LP for the Steiner network problem can be solved to any given precision using efficient combinatorial methods [19], and hence a rounding approach that works with *any* feasible primal solution would yield an efficient and combinatorial $(2+\epsilon)$-approximation for the problem. A second motivation is to determine whether the f-connected subgraph relaxation for the much larger class of skew supermodular functions f possesses the 2-approximate IDP. Our main result, Theorem 1.5, provides some evidence that this may hold. Theorem 1.5 is a decomposition theorem and rounding algorithm that applies to some of the more general f-connector problems studied in [21].

1.2.1. Steiner networks and supermodular functions: Results and terminology. Let $G=(V,E)$ be an undirected graph. A family $\mathcal{F}$ of subsets of V is *skew crossing* if for each $A,B\in\mathcal{F}$, either $A-B, B-A\in\mathcal{F}$ or $A\cap B, A\cup B\in\mathcal{F}$ (or both). Let $f:\mathcal{F}\rightarrow\mathcal{Z}^+$ be an integer-valued function. We call a subgraph H of G *f-connected* if for each $A\in\mathcal{F}$, we have that $|\delta_H(A)|\geq f(A)$. The main problem considered in this paper is that of finding a minimum cost f-connected subgraph for some interesting classes of functions f that capture natural network design problems. We present our arguments as though $\mathcal{F}=\mathcal{P}(V)$, but one easily verifies that the results hold even in the case where f's domain is an arbitrary skew-crossing set family.

We focus on the natural LP relaxation for this problem:

$$P(G,f)=\{x\in[0,1]^E : x(\delta(A))\geq f(A) \text{ for each subset } A\in\mathcal{F}\}. \tag{1}$$

The f-connectivity problem asks us to find an integer vector $x\in P(G,f)$ which minimizes $w\cdot x=\sum_e w_e x_e$. The most general class of functions we consider are *skew supermodular* functions [16] (also called *weakly supermodular* in [25]).[1] A function f is skew supermodular if for each pair of sets $A,B\in\mathcal{F}$, at least one of the following holds:

1. $A\cap B, A\cup B\in\mathcal{F}$ and $f(A)+f(B)\leq f(A\cap B)+f(A\cup B)$,
2. $A-B, B-A\in\mathcal{F}$ and $f(A)+f(B)\leq f(B-A)+f(A-B)$.

The class of $\{0,1\}$ skew supermodular f-connectivity problems captures a variety of well-known combinatorial problems, many of which are outlined in the survey [21]. Let us reconsider a few special cases of this problem.

First if $f(A)=1$ for every proper subset A of V, then this coincides with the minimum spanning tree problem. Given a set of terminals $T\subset V$ if we define f by $f(A)=1$ if A splits T (that is, $A\cap T\neq\emptyset$ and $A\cap T\neq T$), then f captures the NP-hard

[1] Andras Frank, at a workshop in Bertinoro, Italy, has convinced the authors that skew supermodular is a more appropriate name than weakly supermodular. He indicates that David Shmoys suggested this name in 1993.

Steiner tree problem. If $f(A) = 1$ for each $A = \{v\}$ and $f(A) = 0$ otherwise, then this is just the minimum node cover problem. Another case of interest is obtained as follows: consider some pair of nodes $s, t \in V$. Define the following skew supermodular function f: $f(A) = 1$ for each subset A that separates s and t. Then f-connectivity is just asking for the minimum cost s-t path. Suppose that the maximum number of edge-disjoint s-t paths is k, and define $f(A) = 1$ for each A that induces a minimum s-t cut. Then the f-connectivity problem asks for a minimum cost subset of edges which, if we duplicate, increases the connectivity from k to $k + 1$. Call this the s-t connectivity augmentation problem.

For certain classes of functions f, the polytope $P(G, f)$ has integral extreme points. Examples include the shortest path and connectivity augmentation functions defined above. This is not always the case, for instance, the NP-hard Steiner tree problem. It was shown in [36] that for all $\{0, 1\}$ skew supermodular functions the optimum over $P(G, f)$ is no better than a factor of 2 from the optimum over the integer hull of $P(G, f)$. This is proved via a primal-dual algorithm. Jain [25] generalized this to all integer-valued skew supermodular functions using a different approach of iterative rounding.

Encouraged by the decomposition results for Steiner forests in Lemma 1.3, we conjecture the following.

CONJECTURE 1.4. *For any graph G and $\{0, 1\}$ skew supermodular f, if $x \in P(G, f)$ and kx is integral, then there exist f-connected integer vectors $h_1, h_2, \ldots, h_k$ such that $2kx \geq \sum_i h_i$.*

Indeed, we know of no reason why the statement could not hold for general integer-valued skew supermodular functions. Although we are unable to prove the above conjecture, our main theorem establishes some positive evidence by establishing it for certain classes of skew supermodular functions introduced by Goemans and Williamson [20]. We introduce these classes now.

A $\{0, 1\}$ function is termed *maximal* if the following holds: for any disjoint subsets $A, B \subseteq V$, $f(A \cup B) \leq \max(f(A), f(B))$. Equivalently, if A and B are disjoint, then $f(A) = f(B) = 0$ implies that $f(A \cup B) = 0$. A function is *symmetric* if for each $A \subset V$, $f(A) = f(V - A)$. A $\{0, 1\}$ function is *proper* if it is maximal, symmetric, and $f(V) = 0$. Another special class of skew supermodular functions are *downward monotone functions* which satisfy the property that $f(A) \geq f(B)$ if $A \subset B$.

THEOREM 1.5. *For any graph G and $\{0, 1\}$ function f where f is either proper or downward monotone, if $x \in P(G, f)$ and kx is integral, then there exist f-connected integer vectors $h_1, h_2, \ldots, h_k$ such that $2kx \geq \sum_{i=1}^{k} h_i$. Moreover, given x we may find this decomposition in polynomial time.*

Note that the above theorem generalizes Lemma 1.3 since the Steiner forest problem is defined by a $\{0, 1\}$ proper function. One consequence of the above theorem is a new polynomial-time 2-approximation algorithm for the minimum cost f-connectivity problem for proper and downward monotone functions. These new algorithms are primarily of theoretical interest since their running times are not competitive with the primal-dual algorithms [20].

In addition to proving Lemma 1.3 and Theorem 1.5, we consider the general question of whether f-connectivity problems arise in a natural way from other basic problems. We show in section 5 that this is indeed the case for intersecting supermodular functions: they are effectively disguised connectivity augmentation problems in both the directed and undirected settings. We also characterize the (supermodular) functions which define Steiner forest problems. Negative results are given, however, for proper and skew supermodular functions.

1.2.2. Further related work. Our approach to finding approximate integer decompositions for $\{0,1\}$ network design problems amounts to packing forests, each satisfying some connectivity requirement. As alluded to earlier, one special case of this has been considered more extensively in the literature. Given an undirected graph $G = (V, E)$ and set $S \subseteq V$ of terminals, find the maximum number of edge-disjoint S-Steiner trees in G. This problem has been studied from a polyhedral and computational point of view by Grötschel, Martin, and Weismantel [23, 24]. Their motivating application is routing in VLSI design. Kriesell [27] considered the same problem and conjectured that Corollary 1.2 generalizes to packing Steiner trees; that is, if a set S is $2k$–edge-connected in G, then there are k edge-disjoint S-Steiner trees in G. As mentioned earlier, if G is *Eulerian*, Frank, Kiraly, and Kriesell [17] show that if S is $2k$–edge-connected in G, then there are such disjoint Steiner trees. They also showed that if S is $3k$–edge-connected and $V - S$ is a stable set, then there are k edge-disjoint S-Steiner trees. In general graphs, Jain, Mahdian, and Salvatipour [26] showed that if S is k–edge-connected, then there are $\alpha_{|S|}k$ edge-disjoint Steiner trees, where $\alpha_{|S|} \to \frac{4}{|S|}$. They also give results on fractional packing of Steiner trees, and for this case they use the duality between fractional packing and approximation algorithms [9, 22]. As observed in [27, 26], the known results had not guaranteed two edge-disjoint Steiner trees even if S is k–edge-connected for any $k = o(n)$. Recently, Lau [28] showed that a k-packing of Steiner trees can in fact be found if S is $26k$–edge-connected, in the process obtaining the first constant factor approximation for integer packing of Steiner trees. In [29] Lau extended his ideas to the Steiner forest packing problem; given node pairs $s_1t_1, \ldots, s_\ell t_\ell$ such that $\lambda_G(s_it_i) \geq 32k$ for $1 \leq i \leq \ell$ then there are k edge-disjoint forests such that each s_it_i is connected in each of the k forests. This extension was partly motivated by our work in this paper.

2. Preliminaries. The central tool used in this article is that of *splitting off* edges. We state Mader's splitting-off theorem, that was conjectured earlier by Lovász [30].

THEOREM 2.1 (Mader [32]). *Let $G = (V \cup \{s\}, E)$ be an undirected multigraph, where s has positive even degree and s is not incident with a cut edge of G. Then s has two neighbors u and v such that the graph G' obtained from G by replacing su and sv by uv satisfies $\lambda_{G'}(x, y) = \lambda_G(x, y)$ for all $x, y \in V \setminus \{s\}$.*

In this paper we apply the above splitting-off theorem only for Eulerian graphs which do not have cut edges.

The following claim is standard.

CLAIM 2.2. *In an undirected graph G, for any three distinct nodes u, v, w, $\lambda_G(u, w) \geq \min\{\lambda_G(u, v), \lambda_G(v, w)\}$.*

Let S be a proper subset of the nodes V of an undirected graph $G = (V, E)$. We denote by $\delta(S)$ the *cut* induced by S, that is, the subset of edges E with exactly one endpoint in S. For an edge vector $x : E \to \mathbf{R}$ and $E' \subseteq E$, we use $x(E')$ to denote the quantity $\sum_{e \in E'} x_e$. We say that a set X *splits* a set S, or is *S-splitting*, if both $X \cap S$ and $X - S$ are nonempty. We call a set of nodes S an *ℓ-island*, if for any $u, v \in S$, $\lambda_G(u, v) \geq \ell$. From Claim 2.2, it follows that the maximal ℓ-islands are unique and disjoint. We also refer to S as being a *fractional ℓ-island* with respect to some edge vector x^*, if for any S-splitting set U, $x^*(\delta(U)) \geq \ell$.

For a vector $x \in \mathbf{R}_+^E$ we call a subset S' *deficient* if $x(\delta(S')) \leq 1$ and strongly deficient if the inequality is strict. Each *strongly* deficient set S' evidently satisfies $f(S') = 0$ if $x \in P(G, f)$. We make repeated use of the following lemma. It follows

directly from the well-known fact that the function $\delta(S)$ is posi-modular,[2] a definition introduced by Nagamochi and Ibaraki [33].

LEMMA 2.3. *For any graph G and $x \in \mathbf{R}_+^E$, if S' and S'' are deficient sets, then at least one of $S' - S''$ and $S'' - S'$ is deficient.*

We need another simple lemma given below.

LEMMA 2.4. *Let $x^* \in \mathbf{R}_+^E$, and let K be a minimal deficient set. Then K induces a fractional 1-island in the graph obtained by contracting $V - K$ to a single node.*

Proof. Let G^* be obtained by contracting $V - K$ to a single node v^*. If K is not a fractional 1-island, then there exists some proper subset Y' of K such that $x^*(\delta_{G^*}(Y')) = x^*(\delta_G(Y')) < 1$. But then Y' is strongly deficient for x^*, contradicting the minimality of K. □

We give a corollary of Theorem 1.1 that is useful in subsequent sections.

LEMMA 2.5. *Let $G = (V' \cup \{s\}, E)$ be such that V' is a $2k$-island in G and $|\delta_G(s)| \leq 2k$. Then the subgraph induced by V' has k edge-disjoint spanning trees.*

Proof. Let $G' = (V', E')$ be the subgraph of G induced by V'. Let $V_1, V_2, \ldots, V_\ell$ be any partition of V' in G'. We claim that $\sum_{i=1}^{\ell} |\delta_{G'}(V_i)| \geq 2k\ell - |\delta_G(s)| \geq 2k\ell - 2k$, and hence the number of edges in G' between nodes of the partition is at least $k(\ell - 1)$. Thus G' satisfies the conditions of Theorem 1.1 and hence has k edge-disjoint spanning trees. □

3. Packing Steiner forests. In this section we prove Lemma 1.3. Recall that we are given a Eulerian graph $G = (V, E)$ and pairs of nodes $s_1t_1, s_2t_2, \ldots, s_kt_k$ such that $\lambda_G(s_i, t_i) \geq 2k$ for $1 \leq i \leq k$. Given G let $S_1, S_2, \ldots, S_h$ be the maximal $2k$-islands. In fact we prove the following theorem, which can be easily seen to imply Lemma 1.3.

THEOREM 3.1. *Let $G = (V, E)$ be a Eulerian graph, and let $S_1, S_2, \ldots, S_h$ be the maximal $2k$-islands in G. Then, there are k edge-disjoint forests $F_1, F_2, \ldots, F_k$ in G such that, in each F_j and for $1 \leq i \leq h$, S_i is contained in a connected component of F_j. Given G and k, there is an algorithm that finds such a packing in time polynomial in n and $\log k$.*

Proof. The proof is by induction on $|V|$. The base cases with $|V| \leq 2$ are easy to see. We call $v \in V$ a *Steiner node* if v is a singleton island; otherwise it is a *terminal*. We reduce the problem to *basic* instances, defined as instances in which $\cup_j S_j = V$ and $|S_i| \geq 2$ for $1 \leq i \leq h$; in other words there are no Steiner nodes. We get rid of Steiner nodes by splitting off the edges incident to them. Let s be a Steiner node. Since G is Eulerian, $d(s)$ is even. From Theorem 2.1, there are edges su and sv incident to s such that su and sv can be split off without affecting the connectivity of any pair of nodes not involving s. A solution to the problem on the modified graph can be extended to a solution to the original graph by replacing the edge uv by the path consisting of su and sv. Hence we can repeatedly split off edges incident to s until the degree of s is 0. We can eliminate all Steiner nodes in this way and reduce the graph G to a basic instance.

We now assume that G is basic. If $h = 1$, then, from Corollary 1.2, we can find k spanning trees, and hence we are done. If $h \geq 2$, we may apply Lemma 2.4 to find a set $K := \cup_{i \in I} S_i$ such that $|\delta_G(K)| < 2k$, and contracting $V - K$ to a single node s produces a graph G', where K is a $2k$-island. To see this, simply choose x^* to be the edge vector with weight $1/2k$ on each edge and let K be a minimal deficient set. Since a deficient set cannot be S_i-splitting for any i, the S_i's inside K form our set I. From

[2] A function $f : \mathcal{F} \to \mathbf{R}_+$ is posi-modular if $f(A) + f(B) \geq f(A - B) + f(B - A)$ for all $A, B \in \mathcal{F}$.

Lemma 2.5 we can find k edge-disjoint trees in $G[K]$ that do not use edges incident to s. Let these be $T_1, T_2, \ldots, T_k$. Now consider the graph G'' obtained by shrinking K in G to a single node s'. We can apply induction to G'' since it has fewer nodes than G (note that $|K| \geq 2$ since the instance is basic) to obtain edge-disjoint forests $F'_1, F'_2, \ldots, F'_k$ such that each S_i, $i \notin I$, is contained in a single component in each of the forests. We obtain the desired forests $F_1, F_2, \ldots, F_k$ in G as follows: to obtain F_i we replace s' in F'_i with T_i. Note that two nodes u and v which are connected in F'_i via s' will still be connected in F_i since T_i is spanning on K. This finishes the proof of the existence of the packing.

We now prove that the packing can be found in time polynomial in n and $\log k$. To obtain a time polynomial in $\log k$, the decomposition will be output in a compact form with some forests having integer multiplicities. We assume without loss of generality that the number of edges between any two pairs of nodes is at most $2k$; otherwise we can remove some edges without violating the connectivity requirements. We first observe that the maximum number of edge-disjoint spanning trees in a given graph can be found in time polynomial in n and $\log k$ (see Chapter 51, pp. 887–889 in [35]). Therefore the trees in Lemma 2.5 can found in $\text{poly}(n, \log k)$. There are two nontrivial steps to verify polynomial running time.

First, we describe the implementation of the splitting-off step. Let v be a Steiner node with $v_1, v_2, \ldots, v_\ell$ as its neighbors, and let $c(v, v_i)$ be the number of edges between v and v_i. Let $c(v) = \sum_i c(v, v_i)$. From Theorem 2.1, we can split off edges incident to v in pairs. After we split off all of the edges incident to v, let $c'(v_i, v_j)$ be the number of new edges generated between v_i and v_j. It follows that there exists a pair v_i, v_j such that $c'(v_i, v_j) \geq \max\{1, c(v)/(2\ell^2)\}$. For each pair v_i, v_j we can find the maximum number of edges that can be split off at v to generate edges between v_i and v_j by doing a binary search in the range $[0, \min\{c(v, v_i), c(v, v_j)\}]$. Each search involves finding the edge connectivity between all pairs of nodes to ensure that the splitting-off is legal. Thus we can split off edges incident to v in time polynomial in $\log k$ and n.

Second, when G is basic and $h \geq 2$, we need to find a minimal set K such that $|\delta_G(K)| < 2k$. This can be accomplished in polynomial time as follows: we compute the minimum-cut value $\lambda_G(s, t)$ for all node pairs s, t. Pick an arbitrary node u, and let K be the set of all nodes v such that $\lambda_G(u, v) \geq 2k$. From Claim 2.2 it is easy to see that K is a desired minimal set.

This finishes the proof. □

4. Skew supermodular functions. We have essentially examined the problem of decomposing fractional solutions into forests so that any pair of nodes that were originally 1-connected (fractionally) are in the same component of each forest. In this section we study this scheme for more general $\{0, 1\}$ connectivity functions. In particular we prove Theorem 1.5 on proper and downward monotone functions. For skew supermodular functions we describe a reduction to a special case.

4.1. Proper functions. First, we ask if for such a function f, the following property holds: *for any fractional solution to the f-connector problem, a feasible integral solution is obtained from any forest which includes each maximal island in a common component.* We show that that this holds true for the class of $\{0, 1\}$ proper set functions. Hence Lemma 1.3 will imply our desired decomposition result.

THEOREM 4.1. *Let x be a fractional f-connector of $G = (V, E)$, and let $X_1, \ldots, X_l$ be the maximal islands for x. If f is proper, then any forest F that includes each X_i in a common connected component is an f-connector of G.*

Proof. Note that the X_i's partition $V(G)$. It is sufficient to show that $f(S) = 1$ implies that there is an i such that S splits X_i. Suppose this is not the case, and let S be a minimal such set. We may write S as the union of some of the islands. But then by repeated application of the maximality of f, at least one of these islands X must have $f(X) = 1$. Thus by the minimality of S, and without loss of generality, we may assume that $S = X_1$. Now, since X_1 is an island, we have that for each node $u \in X_1$ and each node $v \notin X_1$, there is a subset $S' \subseteq V - X_1$ containing v such that $x(\delta(S')) < 1$; that is, S' is strongly deficient. Note that $f(S') = 0$.

By Lemma 2.3, if S', S'' are strongly deficient sets and $S' - S'', S'' - S' \neq \emptyset$, then at least one of $S' - S'', S'' - S'$ is strongly deficient. Now to complete the proof, consider a minimal collection of strongly deficient sets that covers $V - S$; such a collection exists since each $v \in V - S$ is in a strongly deficient set, as we argued above (recall that $S = X_1$). If for some pair S', S'' we have that $S' - S'', S'' - S'$, and $S' \cap S''$ are nonempty, then by our previous claim, we may assume that $S' - S''$ is also deficient. We may thus replace S' by the set $S' - S''$. Clearly we may repeat this process until the family of strongly deficient sets we obtain is a partition of $V - S$. By our assumption, $f(S) = 1$ and therefore, by symmetricity, $f(V - S) = 1$. However, $V - S$ is the disjoint union of sets S' with $f(S') = 0$ contradicting the maximality of f. This contradiction completes the proof. □

Given a fractional solution $x \in P(G, f)$ let k be such that kx is integral. It follows that $2kx$ induces a Eulerian graph G^*. It is easy to see that the 1-islands induced by x in G are precisely the $2k$-islands in G^*. From Theorem 3.1 in G^* there are k edge-disjoint forests $F_1, F_2, \ldots, F_k$ such that each island is connected in each of the F_i. Thus, from Theorem 4.1 each F_i is an f-connector. This establishes that $2kx$ can be decomposed into k f-connectors when f is a $\{0, 1\}$ proper function. Further the decomposition can be found in time polynomial in n and $\log k$ as shown by Theorem 3.1.

4.2. Downward monotone functions. We now consider the the class of $\{0, 1\}$ downward monotone functions. Recall that f downward monotone implies that $f(A) \geq f(B)$ if $A \subset B$. Whereas for proper functions, one can apply the forest-packing lemma directly, one must do more work in the case of downward monotone functions. We identify a collection of subproblems for which we apply Lemma 2.5, and collectively these will give the desired f-connected forests. Thus the second claim of Theorem 1.5 will be established.

Let $x \in P(G, f)$ and k be an integer such that kx is integral. We denote by G^* the Eulerian graph induced by $2kx$. Suppose there is no strongly deficient set in G. Then G^* has k edge-disjoint spanning trees, each of which is an f-connector, and we are done. Otherwise let $\mathcal{S} = \{S_1, S_2 \ldots, S_\ell\}$ be the minimal strongly deficient sets for x. Lemma 2.3 implies that these sets are disjoint. Let $S = V \setminus (\cup_i S_i)$. Note that S could be the empty set. We observe some useful properties. First, (i) for each i, $f(S_i) = 0$ since S_i is strongly deficient; (ii) by downward monotonicity, $f(Y) = 0$ if $S_i \subseteq Y$. Second, if $S \neq \emptyset$, for any $Y \subseteq S$, Y is not strictly deficient; otherwise $\mathcal{S}$ would not be the set of all minimal strongly deficient sets. Each S_i is a minimal strongly deficient set, and hence from Lemmas 2.4 and 2.5 we can find k disjoint spanning trees in $G^*[S_i]$. Let $\mathcal{T}_i = \{T_{i,1}, \ldots, T_{i,k}\}$ be such a set of trees.

First, we consider the case that $S = \emptyset$, which implies that $S_1, S_2, \ldots, S_\ell$ partition V. Let $\mathcal{F} = \{F_1, \ldots, F_k\}$ be a collection of k edge-disjoint forests, where $F_j = \cup_{i=1}^{\ell} E(T_{i,j})$. It is easy to see that the F_j are edge-disjoint. Each F_j is an f-connector by remark (ii) above.

We now consider the case that $S \neq \emptyset$. Obtain a graph G_1^* from G^* by shrinking $V \setminus S$ into a single node s. We note that S is a $2k$-island in G_1^* since neither S nor any of its subsets was strongly deficient in G. Therefore, for any $u, v \in S$, $\lambda_{G_1^*}(u, v) \geq 2k$. Let the degree of s in G_1^* be $2k'$. Since S was not strongly deficient in G, $k' \geq k$. We modify G_1^* by splitting off edges incident to s, while preserving the connectivity of nodes in S, until the degree of s is exactly $2k$. Let G_2^* be the resulting graph. Using Lemma 2.5, there are k edge-disjoint spanning trees $T_1, \ldots, T_k$ in $G_2^*[S]$. Let E_i be the edge set of T_i. An edge $e \in E_i$ is either an original edge from G^* or is an edge that is obtained by splitting off two edges e', e'' incident to s. In the latter case, note that e' and e'' also correspond to original edges from G^* (possibly incident to distinct nodes in $\cup_i S_i$). Let $E_i' \subseteq E(G^*)$ be the set of edges obtained from E_i by replacing each split-off edge $e \in E_i$ by its corresponding edges e', e''. We remark that E_i' may no longer induce a connected component on S in the graph G^*. Finally, let $e_1, e_2, \ldots, e_{2k}$ be the edges incident to s in G_2^*. We also associate these edges to their original edges in G^*. We obtain the desired edge-disjoint f-connectors $\mathcal{F} = \{F_1, \ldots, F_k\}$ in G^* as follows: we set $F_j = \{e_j\} \cup E_j' \cup (\cup_{i=1}^{\ell} E(T_{i,j}))$. By construction, the F_j's are edge-disjoint.

LEMMA 4.2. *For $1 \leq j \leq k$, F_j is an f-connector.*

Proof. Recall that we already argued the the case when $S = \emptyset$. Let $Y \subset V$ such that $f(Y) = 1$. Note that Y cannot contain S_i for any i; otherwise $f(Y) = 0$ since $f(S_i) = 0$. In addition, if Y "properly" intersects some S_i, then there is an edge $e \in E(T_{i,j})$ that crosses Y (that is, $e \in \delta_{G^*}(Y)$). Therefore it is sufficient to restrict attention to those sets Y such that $Y \subseteq S$. Note that $e_j \in \delta_{G^*}(S)$ and $e_j \in F_j$; therefore e_j satisfies S if $f(S) = 1$. So suppose Y is a proper subset of S. Since T_j is a spanning tree in $G_2^*[S]$, there is an edge e in E_j that crosses Y. If e is an edge from G^*, then $e \in E_j'$ and hence $e \in F_j$. Otherwise e is an edge obtained in the splitting-off process at s, and we replace e by e' and e'' in E_j'. Since at least one of e' and e'' crosses Y in G^*, the proof is complete. □

We have thus shown the existence of k f-connectors in G^*. It remains to argue that these f-connectors can be found in time polynomial in n and $\log k$. We observe that the only nontrivial parts in converting the existence proof into an algorithmic proof are the splitting-off step at s when $S \neq \emptyset$ and the use of Lemma 2.5 to find spanning trees in $G^*[S_i]$, for $1 \leq i \leq \ell$, and in $G_2^*[S]$. The arguments in the proof of Theorem 3.1 can be used identically here to implement these steps in polynomial time.

4.3. Reduction to split instances. We now consider arbitrary $\{0, 1\}$ skew supermodular functions. We describe a reduction of Conjecture 1.4 to a restricted class of instances that we next define. Given a function f and a feasible fractional solution $x \in P(G, f)$, we call (G, f, x) a *split instance* if $x \in P(G, f)$ and there is a subset of nodes $S \subset V$ such that

- for every $A \subseteq S$, $x(\delta(A)) \geq 1$; that is, no subset of S is strongly deficient for x, and
- for every $A \subset V \setminus S$, $f(A) = 0$.

THEOREM 4.3. *Let f be a $\{0, 1\}$ skew supermodular function and $x \in P(G, f)$ such that kx is integral. Given G, x, and k there is an algorithm that obtains a split instance (G', f', x') such that* (i) *f' is a $\{0, 1\}$ skew supermodular function,* (ii) *$x' \in P(G', f')$, and* (iii) *$2kx$ is decomposable into k f-connectors in G if $2kx'$ is decomposable into k f'-connectors in G'.*

If the following conjecture is true, so is Conjecture 1.4.

CONJECTURE 4.4. *Let f be $\{0,1\}$ skew supermodular function on G, and let (G,f,x) induce a split instance. If kx is integral, then there exist f-connected integer vectors $h_1,h_2,\ldots,h_k$ such that $2kx \geq \sum_i h_i$.*

Theorem 4.3 does not claim polynomial time for the algorithm that reduces a given instance to a split instance. We give a sketch of the proof of Theorem 4.3. In the following we assume that kx is integral and that G^* is the Eulerian graph induced by $2kx$ and G. The algorithm starts with an instance (G,f,x) and loops between two phases: a *deficient shrinking phase* and a 1*-set shrinking phase*; and it stops once it produces a split instance. In each phase some nonsingleton subset of nodes Y is shrunk into a single node y, and the connectivity function is modified for the new graph G'. More precisely if f is the original function on G, then we obtain a new $\{0,1\}$ function f' in G' as follows: (i) $f'(\{y\}) = f(Y)$, (ii) for $A \subset V \setminus Y$, $f'(A) = f(A)$, and (iii) for $A \supset Y$, $f'(\{y\} \cup (A \setminus Y)) = f(A)$. It is easy to check that for any Y the function f' is skew supermodular if f is.

Deficient shrinking phase. This phase is similar to the first step in section 4.2 on downward monotone functions. Given $x \in P(G,f)$, let $\mathcal{S} = \{S_1, S_2 \ldots, S_\ell\}$ be the minimal strongly deficient sets for x. Lemma 2.3 implies that these sets are disjoint. Also, by Lemma 2.5, for $1 \leq i \leq \ell$, we can find k edge-disjoint spanning trees in the graph $G_i^* = G^*[S_i]$.

Consider the problem G', f' obtained by shrinking each S_i to a single node and defining f' as the restriction of f to this modified graph. Let $E' \subseteq E(G')$ and E'' be a subset of edges from the G_i^*'s. If E' induces an f'-connected graph in the smaller instance G', f', and E'' includes a spanning tree for each G_i^*, then $E' \cup E''$ induces an f-connected subgraph of G. Moreover, for any f-connected subgraph $H = (V,F)$ we must have that $F \cap E(G')$ induces an f'-connected graph in G'. Thus it suffices to focus on the reduced problem for G', f'.

If we have a split instance after the deficit shrinking step, we stop the procedure. Otherwise, we continue to a 1-set shrinking phase.

1-set shrinking phase. Such a phase begins with an instance G', f' and a subset S (possibly empty and arising from the deficient shrinking phase in G, f, with $S = V(G) - (\cup_i S_i)$) such that (1) for each $v \in V(G') - S$, we have $f'(v) = 0$ and (2) for each subset $Y \subseteq S$, Y is not strongly deficient. Note that property (1) follows from our processing because $f(S_i) = 0$ since S_i was strongly deficient and property (2) follows since otherwise Y would contain a minimal strongly deficient set and so would have been one of the S_i's in the deficient shrinking phase.

We now consider any minimal $A \subseteq V(G') \setminus S$ such that $f'(A) = 1$. One notes that the minimal sets of this type are node-disjoint by the skew supermodularity of f'. Also, if there is no such set, we have a split instance, and so we would have terminated. Note also that since each $v \in V(G') \setminus S$ satisfies $f'(v) = 0$, we have that $|A| \geq 2$ for any such A. We shrink A without affecting feasibility, since any set Y with $f'(Y) = 1$ is not A-splitting; for otherwise skew supermodularity would imply a proper subset of A has $f'(A) = 1$ which contradicts the minimality of A. Since $|A| \geq 2$, such a shrinking operation reduces the size of the graph. Upon completion of 1-set shrinking we return to deficient shrinking.

This completes the description of the procedure. After every pair of phases, we shrink some nontrivial subset, and hence after at most n iterations, we obtain a split instance. This finishes the proof sketch of Theorem 4.3.

Recall that we do not claim that the the reduction to a split instance can be carried out in polynomial time. The bottleneck is the step in the 1-set reduc-

tion that requires us to find a minimal set $A \subseteq V(G') \setminus S$ such that $f'(A) = 1$.

5. What problem is f-connectivity solving? Given a specific $\{0,1\}$ skew supermodular function f, it is natural to ask, *What problem is f-connectivity solving*? In other words, in which cases does a supermodular function f encode a problem of more natural combinatorial significance? To make this more concrete, we give several (positive and negative) results related to this agenda.

In each case, we may have a graph $G = (V, E)$ and a skew supermodular function f. Our goal is to build minimum cost networks that are f-connected. Throughout this section, we refer to a set A as *good* if $f(A) = 1$; otherwise it is *bad*. As usual, G represents where we may install capacity, and so it does not play a central role in this section. Instead, we explore whether certain functions f can be interpreted as a connectivity problem in a related network.

5.1. Connectivity augmentation. We first examine two instances where f-connectivity is encoding a *connectivity augmentation problem* in some graph $G' = (V, E')$, in other words, where there is a set of edges E' and list of terminal pairs $s_i t_i$ for some $i = 1, 2 \ldots, k$ such that a set $S \subseteq V$ is good (for f) if and only if $\delta_{G'}(S) = \emptyset$ and there is some pair with $s_i \in S, t_i \notin S$.

5.1.1. Fully supermodular functions. A set function f is *fully supermodular* if $f(V) = f(\emptyset) = 0$ and for all A, B we have $f(A)+f(B) \leq f(A \cup B)+f(A \cap B)$. Again, more generally these may be defined in terms of an intersecting family $\mathcal{F}$ of sets, but we only present our arguments in the case where all sets are in the family. Such functions can be used to generalize a number of classical results in combinatorial optimization, including Edmonds' disjoint branching theorem [11]. This was proposed by Frank [15], who actually introduced the more general class of *intersecting supermodular* functions that only require the inequality above for A, B, with nonempty intersection.

We show that f-connectivity network design for $\{0,1\}$ fully supermodular functions f arises as a *connectivity augmentation problem*. Namely, we show that there is a set of "auxiliary" edges E' such that in the graph $G' = (V, E')$ there exist nodes s, t and $f(S) = 1$ if and only if S is $\{s, t\}$-splitting and $\delta(S) \cap E' = \emptyset$. Thus finding a minimum cost f-connected graph is the same as finding a minimum cost set $F \subseteq E$ of edges such that s, t are connected in $G[V, E' \cup F]$. We mention that the following argument applies equally well to directed network design problems.

For any pair of good sets A, B we have $A \cap B, A \cup B$ are also good. Thus there is a unique maximal good set M and a unique minimal good set S. Since $M \neq V, S \neq \emptyset$, we may choose an arbitrary $s \in S$ and an arbitrary $t \in V - M$, and so every good set contains s and not t.

We obtain the above claimed G' by starting with the empty graph $G^0 := (V, \emptyset)$ and adding edges in an iterative fashion. In iteration i we find a single edge e^i that we add to G^i to obtain G^{i+1}. This is done as follows: suppose there is some bad set X such that $\delta_{G^i}(X) = \emptyset$. (If there is no such set, then $G' = G^i$, and the procedure terminates.) We show that there is some edge $e^i = uv$ such that $u \in X, v \notin X$, and e^i is not contained in any $\delta(A)$ for a good set A. Suppose this is not the case; then for each $u \in X$ and $v \in V - X$, there is a good set Y_{uv} containing u but not v. Fix some $v \in V - X$, and note that $Y(v) = \cup_{u \in X} Y_{uv}$ is also good and $X \subseteq Y(v)$. But then $\cap_{v \in V - X} Y(v)$ is also good, and evidently this set is just X, a contradiction. Thus after some $\ell \leq \binom{n}{2}$ iterations, we have that in G^ℓ a cut $\delta_{G^\ell}(A)$ is empty if and only if it is good. A similar proof yields an analogous result for directed graphs.

5.1.2. Fastidious functions. Recall that a $\{0,1\}$ proper function f is a symmetric set function $f : V \to \{0,1\}$ such that $f(A \cup B) \leq \max\{f(A), f(B)\}$. In the next section, we see that not all proper functions arise from connectivity augmentation. In this section we consider a subclass of proper functions that arise from Steiner forest problems. A symmetric function is *fastidious* if no good set is the union of bad sets. (For proper functions, no good set is the *disjoint* union of bad sets.) We show that *fastidious functions are precisely those that encode Steiner forest problems.* One direction is clear: a Steiner forest problem obviously gives rise to a fastidious function. We now show the converse.

Given a fastidious f, we define a graph $H = (V, E(H))$, where

$$E(H) = \{uv : \text{every } \{u,v\}\text{-splitting set } S \text{ is good (that is, } f(S) = 1)\}.$$

Let $S_1, S_2, \ldots, S_\ell$ be the connected components of H. We now claim that a set X is good according to f if and only if it splits some S_i, which would give us the desired result. Suppose this is not the case, and let X be a minimal good set such that $\delta_H(X) = \emptyset$. That is, for any $u \in X$ and $v \in V - X$, there is a bad set Y_{uv} such that $u \in Y_{uv}$ and $v \notin Y_{uv}$. Consider two cases. Suppose first that for some pair such a set exists with $X - Y_{uv}$ nonempty. Then, since f is proper, either $X - Y_{uv}$ or $Y_{uv} \cap X$ is good. By minimality of X, there is some edge zw of H with $z \in X - Y_{uv}$ and $w \in Y_{uv} \cap X$. But any such edge must lie in $\delta_H(Y_{uv})$, contradicting the fact that $f(Y_{uv}) = 0$.

In the second case, for every pair uv with $u \in X, v \in V \setminus X$, we have $X \subseteq Y_{uv}$. But then the $(V - Y_{uv})$'s are a collection of bad sets whose union is the good set $V - X$, a contradiction.

5.2. Skew supermodular functions and embedded connectivity. We have seen several cases of supermodular functions encoding an underlying (or hidden) connectivity problem. We cannot expect to be as lucky for general skew supermodular functions. Consider the node cover problem that arises from the skew supermodular function $f : V \to \{0,1\}$, where a set S is good if and only if S is a singleton. One may deduce that there is no graph $G' = (V, E')$ for which f encodes a connectivity augmentation problem on G'. However, we may cast the problem in this form if we embed it in a larger graph and allow higher connectivity requirements: take $H = (V + s, \{sv : v \in V\})$. Consider the problem of adding edges E'' to H so that each node v is 2–edge-connected to s in $H + E''$. The good sets for f now correspond precisely to the "deficit cuts" for this 2-connectivity problem. In general, we may define an *embedded connectivity problem* as consisting of a pair of graphs $G = (V, E), H' = (V', E')$, with $V \subseteq V'$; ℓ node pairs $s_1t_1, \ldots, s_\ell t_\ell$, where $s_i, t_i \in V'$ for each i; and integers k_i for $i = 1, 2 \ldots, \ell$. For such an instance, we call a subset $S \subseteq V'$ a *target set* if for some i, S is $\{s_i, t_i\}$-splitting, and $|\delta_{H'}(S)| < k_i$, the target connectivity for s_i, t_i. The following is easily shown.

FACT 5.1. *Any embedded connectivity problem gives rise to a skew supermodular function.*

We believe that analyzing which functions encode embedded connectivity problems is a potentially fruitful direction for further study. Such embedding problems would seem, however, of limited use unless the size of H' is polynomially bounded in G (and somehow f). The focus should thus be on *p-bounded* problems, where in addition $|V'| \leq p(|V|)$, for some polynomial p.

6. Conclusions. We have shown how splitting off, combined with Theorem 1.1, yields decomposition and rounding algorithms for a large class of 0-1 network design

problems. Several open problems remain. First, it would be interesting to resolve Conjecture 1.4. If the decomposition algorithm can be generalized to integer-valued skew supermodular functions, it would yield an alternative algorithm to that of Jain [25]. It would also yield a combinatorial rounding algorithm for the Steiner network problem. Second, the inverse f-connectivity questions raised in section 5 are of interest in their own right and may also prove useful in resolving Conjecture 1.4.

Acknowledgment. We thank the anonymous referees for useful comments that greatly improved the paper.

REFERENCES

[1] M. Andrews and L. Zhang, *Bounds on fiber minimization in optical networks with fixed fiber capacity*, in Proceedings of the 24th IEEE INFOCOM Conference, Miami, FL, 2005, pp. 409–419.

[2] M. Andrews and L. Zhang, *Complexity of wavelength assignment in optical network optimization*, in Proceedings of the 25th IEEE INFOCOM Conference, Barcelona, Spain, 2006.

[3] A. Agrawal, P. Klein, and R. Ravi, *When trees collide: An approximation algorithm for the generalized Steiner problem on networks*, SIAM J. Comput., 24 (1995), pp. 440–456.

[4] J. Bang-Jensen, A. Frank, and B. Jackson, *Preserving and increasing local edge-connectivity in mixed graphs*, SIAM J. Discrete Math., 8 (1995), pp. 155–178.

[5] S. Baum and L. E. Trotter, Jr., *Integer rounding and polyhedral decomposition for totally unimodular systems*, in Optimization and Operations Research, Lecture Notes in Econ. and Math. Systems 157, R. Henn, B. Korte, and W. Oettli, eds., Springer, Berlin, 1978, pp. 15–23.

[6] S. Baum and L. E. Trotter, Jr., *Integer rounding for polymatroid and branching optimization problems*, SIAM J. Algebraic Discrete Meth., 2 (1981), pp. 416–425.

[7] D. Bienstock, M. X. Goemans, D. Simchi-Levi, and D. Williamson, *A note on the prize collecting traveling salesman problem*, Math. Program., 59 (1993), pp. 413–420.

[8] A. Blum, R. Ravi, and S. Vempala, *A constant-factor approximation for the k-MST problem*, J. Comput. System Sci., 58 (1999), pp. 101–108.

[9] R. Carr and S. Vempala, *Randomized meta-rounding*, Random Structures Algorithms, 20 (2002), pp. 343–352.

[10] C. Chekuri, M. Mydlarz, and F. B. Shepherd, *Multicommodity demand flow in a tree and packing integer programs*, ACM Trans. Algorithms, 3 (2007).

[11] J. Edmonds, *Edge-disjoint branchings*, in Combinatorial Algorithms, B. Rustin, ed., Academic Press, New York, 1973, pp. 91–96.

[12] J. Edmonds and R. Giles, *A min-max relation for submodular functions on graphs*, Ann. Discrete Math., 1 (1977), pp. 185–204.

[13] T. Erlebach, A. Pagourtzis, K. Potika, and S. Stefanakos, *Resource allocation problems in multifiber WDM tree networks*, in Proceedings of Workshop on Graph Theoretic Concepts in Computer Science, Lecture Notes in Comput. Sci. 2880, Springer, Berlin, 2003, pp. 218–229.

[14] T. Fukunaga and H. Nagamochi, *The set connector problem in graphs*, in Proceedings of the 12th IPCO Conference, Ithaca, NY, Mathematical Programming Society, 2007, pp. 484–498.

[15] A. Frank, *Kernel systems of directed graphs*, Acta Sci. Math., 41 (1979), pp. 63–76.

[16] A. Frank, *Applications of relaxed submodularity*, in Proceedings of the International Congress of Mathematicians, Berlin, 1998, Vol. III: Invited Lectures, Documenta Mathematica, Extra Volume ICM 1998, G. Fischer and U. Rehmann, eds., Documenta Mathematica, Berlin, 1998, pp. 343–354.

[17] A. Frank, T. Kiraly, and M. Kriesell, *On decomposing a hypergraph into k connected sub-hypergraphs*, Discrete Appl. Math., 131 (2003), pp. 373–383.

[18] N. Garg, *A 3-approximation for the minimum tree spanning k vertices*, in Proceedings of the 37th IEEE Symposium on Foundations of Computer Science, Burlington, VT, 1996, pp. 302–309.

[19] N. Garg and R. Khandekar, *Fast approximation algorithms for fractional Steiner forest and related problems*, in Proceedings of the 43rd IEEE Symposium on Foundations of Computer Science, Vancouver, Canada, 2002, pp. 500–509.

[20] M. X. Goemans and D. P. Williamson, *A general approximation technique for constrained forest problems*, SIAM J. Comput., 24 (1995), pp. 296–317.
[21] M. Goemans and D. Williamson, *The primal-dual method for approximation algorithms and its application to network design problems*, in Approximation Algorithms for NP-Hard Problems, D. Hochbaum, ed., PWS, Boston, 1997.
[22] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer, Berlin, 1988.
[23] M. Grötschel, A. Martin, and R. Weismantel, *Packing Steiner trees: Polyhedral investigations*, Math. Program., 72 (1996), pp. 101–123.
[24] M. Grötschel, A. Martin, and R. Weismantel, *Packing Steiner trees: Separation algorithms*, SIAM J. Discrete Math., 9 (1996), pp. 233–257.
[25] K. Jain, *A factor* 2 *approximation algorithm for the generalized Steiner network problem*, Combinatorica, 21 (2001), pp. 39–60.
[26] K. Jain, M. Mahdian, and M. R. Salvatipour, *Packing Steiner trees*, in Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms, Baltimore, MD, 2003, pp. 266–274.
[27] M. Kriesell, *Local spanning trees in graphs and hypergraph decomposition with respect to edge connectivity*, Proceedings of the 6th Cologne-Twente Workshop on Graphs and Combinatorial Optimization, Electron. Notes Discrete Math. 3, Elevier, Amsterdam, 1999.
[28] L. C. Lau, *An approximate max-Steiner-tree-packing min-Steiner-cut theorem*, Combinatorica, 27 (2007), pp. 71–90.
[29] L. C. Lau, *Packing Steiner forests*, in Proceedings of the 11th IPCO Conference, Berlin, Mathematical Programming Society, 2005, pp. 362–276.
[30] L. Lovász, *Unsolved Problems*, in Proceedings of the Fifth British Combinatorial Conference, Aberdeen, 1975, Congr. Numer., XV, C. St. J. A. Nash-Williams, Win-Sheehan, eds., Utilitas Mathematica, Winnipeg, Manitoba, 1976, pp. 638–685.
[31] C. Lucchesi and D. Younger, *A minimax theorem for directed graphs*, J. London Math. Soc. (2), 17 (1978), pp. 369–374.
[32] W. Mader, *A reduction method for edge-connectivity in graphs*, in Advances in Graph Theory, Ann. of Discrete Math. 3, B. Bollobas, ed., North–Holland, Amsterdam, 1978, pp. 145–164.
[33] H. Nagamochi and T. Ibaraki, *Polyhedral structure of submodular and posi-modular systems*, Discrete Appl. Math., 107 (2000), pp. 165–189.
[34] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley and Sons, New York, 1987.
[35] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency*, Springer, Berlin, 2003.
[36] D. P. Williamson, M. X. Goemans, M. Mihail, and V. Vazirani, *A primal-dual approximation algorithm for generalized Steiner network problems*, Combinatorica, 15 (1995), pp. 435–454.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 178–194

INTEGRALITY GAPS OF SEMIDEFINITE PROGRAMS FOR VERTEX COVER AND RELATIONS TO ℓ_1 EMBEDDABILITY OF NEGATIVE TYPE METRICS*

HAMED HATAMI†, AVNER MAGEN†, AND EVANGELOS MARKAKIS‡

Abstract. We study various semidefinite programming (SDP) formulations for VERTEX COVER by adding different constraints to the standard formulation. We show that VERTEX COVER cannot be approximated better than $2 - O(\sqrt{\log\log n/\log n})$ even when we add the so-called pentagonal inequality constraints to the standard SDP formulation, and thus almost meet the best upper bound known due to Karakostas [*Proceedings of the 32nd International Colloquium on Automata, Languages and Programming*, 2005], of $2 - \Omega(\sqrt{1/\log n})$. We further show the surprising fact that by strengthening the SDP with the (intractable) requirement that the metric interpretation of the solution embeds into ℓ_1 with no distortion, we get an exact relaxation (integrality gap is 1), and on the other hand, if the solution is arbitrarily close to being ℓ_1 embeddable, the integrality gap is $2 - o(1)$. Finally, inspired by the above findings, we use ideas from the integrality gap construction of Charikar [*SODA* '02: *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, Philadelphia, 2002, pp. 616–620] to provide a family of simple examples for negative type metrics that cannot be embedded into ℓ_1 with distortion better than $8/7 - \epsilon$. To this end we prove a new isoperimetric inequality for the hypercube.

Key words. vertex cover, semidefinite programming, integrality gap

AMS subject classifications. 90C22, 68W25, 54C25

DOI. 10.1137/070700103

1. Introduction.

A VERTEX COVER in a graph $G = (V, E)$ is a set $S \subseteq V$ such that every edge $e \in E$ intersects S in at least one endpoint. Denote by $\mathrm{vc}(G)$ the size of the minimum vertex cover of G. It is well known that the minimum vertex cover problem has a 2-approximation algorithm, and it is widely believed that for every constant $\epsilon > 0$, there is no $(2-\epsilon)$-approximation algorithm for this problem. Currently the best-known hardness result for this problem, based on the PCP theorem, shows that 1.36-approximation is NP-hard [10]. If we were to assume the Unique Games Conjecture [19], the problem would be essentially settled as $2 - \Omega(1)$ would then be NP-hard [20].

In [15], Goemans and Williamson introduced semidefinite programming (SDP) as a tool for obtaining approximation algorithms. Since then semidefinite programming has become an important technique, and for many problems the best-known approximation algorithms are obtained by solving an SDP relaxation of them.

The best-known algorithms for VERTEX COVER compete in "how big is the little oh" in the $2-o(1)$ factor. The best two are in fact based on SDP relaxations: Halperin [16] gives a $(2 - \Omega(\log\log\Delta/\log\Delta))$ approximation where Δ is the maximal degree of the graph while Karakostas obtains a $(2 - \Omega(1/\sqrt{\log n}))$ approximation [18]. As we later show, our lower bound almost meets the latter upper bound even in this resolution of the little oh.

*Received by the editors August 14, 2007; accepted for publication (in revised form) July 2, 2008; published electronically December 17, 2008. A preliminary version of this work appears in [17].
http://www.siam.org/journals/sidma/23-1/70010.html

†Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada (hamed@cs.toronto.edu, avner@cs.toronto.edu).

‡Corresponding author. Center for Math and Computer Science (CWI), Amsterdam, The Netherlands (vangelis@cwi.nl).

The standard way to formulate the VERTEX COVER problem as a quadratic integer program is the following:

$$\begin{array}{lll} \textbf{Min} & \sum_{i\in V}(1+x_0x_i)/2 & \\ \text{s.t.} & (x_i-x_0)(x_j-x_0)=0 & \forall\ ij\in E,\\ & x_i\in\{-1,1\} & \forall\ i\in\{0\}\cup V, \end{array}$$

where the set of the vertices i for which $x_i = x_0$ corresponds to the vertex cover. Relaxing this integer program to a semidefinite program, the scalar variable x_i becomes a vector $\mathbf{v}_i$ and we get

$$\begin{array}{lll} \textbf{Min} & \sum_{i\in V}(1+\mathbf{v}_0\mathbf{v}_i)/2 & \\ \text{s.t.} & (\mathbf{v}_i-\mathbf{v}_0)\cdot(\mathbf{v}_j-\mathbf{v}_0)=0 & \forall\ ij\in E,\\ & \|\mathbf{v}_i\|=1 & \forall\ i\in\{0\}\cup V. \end{array} \tag{1}$$

Kleinberg and Goemans [22] proved that SDP (1) has an integrality gap of $2-o(1)$. Specifically, given $\epsilon > 0$ they construct a graph G_ϵ for which $\mathrm{vc}(G_\epsilon)$ is at least $(2-\epsilon)$ times larger than the solution to SDP (1). They also suggested the following strengthening of SDP (1) and left its integrality gap as an open question:

$$\begin{array}{lll} \textbf{Min} & \sum_{i\in V}(1+\mathbf{v}_0\mathbf{v}_i)/2 & \\ \text{s.t.} & (\mathbf{v}_i-\mathbf{v}_0)\cdot(\mathbf{v}_j-\mathbf{v}_0)=0 & \forall\ ij\in E,\\ & (\mathbf{v}_i-\mathbf{v}_k)\cdot(\mathbf{v}_j-\mathbf{v}_k)\geq 0 & \forall\ i,j,k\in\{0\}\cup V,\\ & \|\mathbf{v}_i\|=1 & \forall\ i\in\{0\}\cup V. \end{array} \tag{2}$$

Charikar [6] answered this question by showing that the same graph G_ϵ but a different vector solution satisfies SDP (2)[1] and gives rise to an integrality gap of $2-o(1)$ as before. The following is an equivalent formulation to SDP (2):

$$\begin{array}{lll} \textbf{Min} & \sum_{i\in V}1-\|\mathbf{v}_0-\mathbf{v}_i\|^2/4 & \\ \text{s.t.} & \|\mathbf{v}_i-\mathbf{v}_0\|^2+\|\mathbf{v}_j-\mathbf{v}_0\|^2=\|\mathbf{v}_i-\mathbf{v}_j\|^2 & \forall\ ij\in E,\\ & \|\mathbf{v}_i-\mathbf{v}_k\|^2+\|\mathbf{v}_j-\mathbf{v}_k\|^2\geq\|\mathbf{v}_i-\mathbf{v}_j\|^2 & \forall\ i,j,k\in\{0\}\cup V,\\ & \|\mathbf{v}_i\|=1 & \forall\ i\in\{0\}\cup V. \end{array} \tag{3}$$

Viewing SDPs as relaxations over ℓ_1. The above reformulation reveals a connection to metric spaces. The second constraint in SDP (3) says that $\|\cdot\|^2$ induces a metric on $\{\mathbf{v}_i : i\in\{0\}\cup V\}$, while the first says that $\mathbf{v}_0$ is on the shortest path between the images of every two neighbors. This suggests a more careful study of the problem from the metric viewpoint, which is the purpose of this article. Such connections are also important in the context of the SPARSEST CUT problem, where the natural SDP relaxation was analyzed in the breakthrough work of Arora, Rao, and Vazirani [5] and it was shown that its integrality gap is at most $O(\sqrt{\log n})$. This later gave rise to some significant progress in the theory of metric spaces [7, 4].

Let $f : (X,d)\to(X',d')$ be an embedding of metric space (X,d) into another metric space (X',d'). The value $\sup_{x,y\in X}\frac{d'(f(x),f(y))}{d(x,y)}\times\sup_{x,y\in X}\frac{d(x,y)}{d'(f(x),f(y))}$ is called the distortion of f. For a metric space (X,d), let $c_1(X,d)$ denote the minimum distortion required to embed (X,d) into ℓ_1. Notice that $c_1(X,d)=1$ if and only if (X,d) can be embedded isometrically into ℓ_1, namely, without changing any of the distances. Consider a vertex cover S and its corresponding solution to SDP (2), i.e.,

[1] To be more precise, Charikar's result was about a slightly weaker formulation than (2) but it is not hard to see that the same construction works for SDP (2) as well.

$\mathbf{v}_i = 1$ for every $i \in S \cup \{0\}$ and $\mathbf{v}_i = -1$ for every $i \notin S$. The metric defined by $\|\cdot\|^2$ on this solution (i.e., $d(i,j) = \|\mathbf{v}_i - \mathbf{v}_j\|^2$) is isometrically embeddable into ℓ_1. Thus we can strengthen SDP (2) by allowing any arbitrary list of valid inequalities in ℓ_1 to be added. The triangle inequality is one type of such constraints. The next natural inequality of this sort is the pentagonal inequality: A metric space (X,d) is said to satisfy the pentagonal inequality if for $S,T \subset X$ of sizes 2 and 3, respectively, it holds that $\sum_{i\in S, j\in T} d(i,j) \geq \sum_{i,j\in S} d(i,j) + \sum_{i,j\in T} d(i,j)$. Note that this inequality does not apply to every metric, but it does hold for those that are ℓ_1-embeddable. This leads to the following natural strengthening of SDP (3):

(4)
$$\begin{array}{lll}
\textbf{Min} & \sum_{i\in V} 1 - \|\mathbf{v}_0 - \mathbf{v}_i\|^2/4 & \\
\text{s.t.} & \|\mathbf{v}_i - \mathbf{v}_0\|^2 + \|\mathbf{v}_j - \mathbf{v}_0\|^2 = \|\mathbf{v}_i - \mathbf{v}_j\|^2 & \forall\ ij \in E, \\
& \sum_{i\in S, j\in T} \|\mathbf{v}_i - \mathbf{v}_j\|^2 \geq \sum_{i,j\in S} \|\mathbf{v}_i - \mathbf{v}_j\|^2 + \sum_{i,j\in T} \|\mathbf{v}_i - \mathbf{v}_j\|^2 & \forall\ S,T \subseteq \{0\}\cup V,\ |S|=2, |T|=3 \\
& \|\mathbf{v}_i\| = 1 & \forall\ i \in \{0\}\cup V.
\end{array}$$

In Theorem 5, we prove that SDP (4) has an integrality gap of $2 - o(1)$. It is important to point out that a priori there is no reason to believe that local addition of inequalities such as these will not improve the integrality gap; indeed in the case of Sparsest Cut triangle inequality is necessary to achieve the $O(\sqrt{\log n})$ bound mentioned above. It is interesting to note that for Sparsest Cut, it is not known how to show a nonconstant integrality gap against pentagonal (or any other k-gonal) inequalities, although recently a nonconstant integrality gap was shown in [21] and later in [8], in the presence of the triangle inequalities.[2]

One can further impose any ℓ_1-constraint not only for the metric defined by $\{\mathbf{v}_i : i \in V \cup \{0\}\}$, but also for the one that comes from $\{\mathbf{v}_i : i \in V \cup \{0\}\} \cup \{-\mathbf{v}_i : i \in V \cup \{0\}\}$. Triangle inequalities for this extended set result in the constraints $\|\mathbf{v}_i - \mathbf{v}_j\|^2 + \|\mathbf{v}_i - \mathbf{v}_k\|^2 + \|\mathbf{v}_j - \mathbf{v}_k\|^2 \leq 2$. The corresponding tighter SDP is used in [18] to get an integrality gap of at most $2 - \Omega(\frac{1}{\sqrt{\log n}})$. Karakostas [18] asks whether the integrality gap of this strengthening breaks the "$2 - o(1)$ barrier": we answer this negatively in section 4.3. In fact we show that the above upper bound is almost asymptotically tight, exhibiting integrality gap of $2 - O(\sqrt{\frac{\log\log n}{\log n}})$.

Integrality gap with respect to ℓ_1 embeddability. At the extreme, strengthening the SDP with ℓ_1-valid constraints would imply the condition that the metric defined by $\|\cdot\|$ on $\{\mathbf{v}_i : i \in \{0\}\cup V\}$, namely, $d(i,j) = \|\mathbf{v}_i - \mathbf{v}_j\|^2$, is ℓ_1 embeddable. Doing so leads to the following intractable program:

(5)
$$\begin{array}{lll}
\textbf{Min} & \sum_{i\in V} 1 - \|\mathbf{v}_0 - \mathbf{v}_i\|^2/4 & \\
\text{s.t.} & \|\mathbf{v}_i - \mathbf{v}_0\|^2 + \|\mathbf{v}_j - \mathbf{v}_0\|^2 = \|\mathbf{v}_i - \mathbf{v}_j\|^2 & \forall\ ij \in E \\
& \|\mathbf{v}_i\| = 1 & \forall\ i \in \{0\}\cup V \\
& c_1(\{\mathbf{v}_i : i \in \{0\}\cup V\}, \|\cdot\|^2) = 1. &
\end{array}$$

In [1], it is shown that an SDP formulation of Minimum Multicut, even with the constraint that the $\|\cdot\|^2$ distance over the variables is isometrically embeddable into ℓ_1, still has a large integrality gap. Let us next consider the Max Cut problem, which is more intimately related to our problem. For this problem it is easy to see

[2] As Khot and Vishnoi note, and leave as an open problem, it is possible that their example satisfies some or all k-gonal inequalities.

that K_3, the complete graph on three vertices, exhibits an integrality gap of 8/9. Now since every metric space on three points is isometrically embeddable into ℓ_1, the ℓ_1 embeddability does not prevent integrality gap of 8/9.[3] It is, therefore, tempting to believe that there is a large integrality gap for SDP (5) as well. Surprisingly, SDP (5) has no gap at all: we show in Theorem 2 that the value of SDP (5) is exactly the size of the minimum vertex cover. A consequence of this fact is that any feasible solution to SDP (2) that surpasses the minimum vertex cover induces an ℓ_2^2 distance which is not isometrically embeddable into ℓ_1. This includes the integrality gap constructions of Kleinberg and Goemans, and that of Charikar's for SDPs (2) and (3), respectively. The construction of Charikar is more interesting in this context as the obtained ℓ_2^2 distance is also a negative type metric, that is, an ℓ_2^2 metric that satisfies triangle inequality. See [9] for background and nomenclature.

In contrast to Theorem 2, we show in Theorem 3 that if we relax the embeddability constraint in SDP (5) to $c_1(\{\mathbf{v}_i : i \in \{0\} \cup V\}, \|\cdot\|^2) \leq 1 + \delta$ for any constant $\delta > 0$, then the integrality gap may "jump" to $2 - o(1)$. Compare this with a problem such as Sparsest Cut in which an addition of such a constraint immediately implies integrality gap at most $1 + \delta$.

Negative type metrics that are not ℓ_1 embeddable. Negative type metrics are metrics which are the squares of Euclidean distances of set of points in Euclidean space. Inspired by Theorem 2, we construct in section 5 a simple negative type metric space $(X, \|\cdot\|^2)$ that does not embed well into ℓ_1. Specifically, we get $c_1(X) \geq \frac{8}{7} - \epsilon$ for every $\epsilon > 0$. In order to show this we prove a new isoperimetric inequality for the hypercube $Q_n = \{-1, 1\}^n$, which we believe is of independent interest. This theorem generalizes the standard one, and under certain conditions provides better guarantees for edge expansion.

Theorem 1 (generalized isoperimetric inequality). *For every set $S \subseteq Q_n$,*

$$|E(S, S^c)| \geq |S|(n - \log_2 |S|) + p(S),$$

where $p(S)$ denotes the number of vertices $\mathbf{u} \in S$ such that $-\mathbf{u} \in S$.

Khot and Vishnoi [21] constructed an example of an n-point negative type metric that for every $\delta > 0$ requires distortion at least $(\log\log n)^{1/6-\delta}$ to embed into ℓ_1. Krauthgamer and Rabani [23] showed that, in fact, Khot and Vishnoi's example requires a distortion of at least $\Omega(\log\log n)$. Later Devanur, Khot, Saket, and Vishnoi [8] showed an example with distortion $\Omega(\log\log n)$ even on average when embedded into ℓ_1 (we note that our example is also "bad" on average). Although the above examples require nonconstant distortion to embed into ℓ_1, we believe that our result is still interesting because (i) our construction is much simpler than the ones in [8, 21, 23]; in comparison, showing that triangle inequality holds requires a lot of technical work in [8, 21, 23], whereas in our construction it is immediate that (ii) very few examples are known of negative type metrics that do not embed isometrically into ℓ_1, and any such example reveals some underlying structure. Prior to Khot and Vishnoi's result, the best-known lower bounds (see [21]) were due to Vempala, 10/9 for a metric obtained by a computer search, and Goemans, 1.024 for a metric based on the Leech Lattice. We mention that by [4] every negative type metric embeds into ℓ_1 with distortion $O(\sqrt{\log n}\log\log n)$.

[3] Notice that an SDP to this problem does not have the auxiliary vector $\mathbf{v}_0$ (as does SDP (5)) in addition to the vectors that correspond to the vertices of the graph, but even if we add such a vector, it has no effect on the program, and it could be simply set to identify with one of the other vectors.

2. Preliminaries and notation. A vertex cover of a graph G is a set of vertices that touch all edges. An independent set in G is a set $I \subseteq V$ such that no edge $e \in E$ joins two vertices in I. We denote by $\alpha(G)$ the size of the maximum independent set of G. Vectors are always denoted in bold font (such as $\mathbf{v}$, $\mathbf{w}$, etc.); $\|\mathbf{v}\|$ stands for the Euclidean norm of $\mathbf{v}$, $\mathbf{u} \cdot \mathbf{v}$ for the inner product of $\mathbf{u}$ and $\mathbf{v}$, and $\mathbf{u} \otimes \mathbf{v}$ for their tensor product. Specifically, if $\mathbf{v}, \mathbf{u} \in \mathbb{R}^n$, $\mathbf{u} \otimes \mathbf{v}$ is the vector with coordinates indexed by ordered pairs $(i,j) \in [n]^2$ that assumes value $\mathbf{u}_i \mathbf{v}_j$ on coordinate (i,j). Similarly, the tensor product of more than two vectors is defined. It is easy to see that $(\mathbf{u} \otimes \mathbf{v}) \cdot (\mathbf{u}' \otimes \mathbf{v}') = (\mathbf{u} \cdot \mathbf{u}')(\mathbf{v} \cdot \mathbf{v}')$. For two vectors $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^m$, denote by $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^{n+m}$ the vector whose projection to the first n coordinates is $\mathbf{u}$ and to the last m coordinates is $\mathbf{v}$.

Next, we give a few basic definitions and facts about finite metric spaces. As we have already defined above, we say that a metric space (X, d_X) embeds with distortion at most D into (Y, d_Y), if there exists a mapping $\phi : X \to Y$ so that for all $a, b \in X$, $\gamma \cdot d_X(a,b) \leq d_Y(\phi(a), \phi(b)) \leq \gamma D \cdot d_X(a,b)$, for some $\gamma > 0$. We say that (X,d) is ℓ_1 embeddable if it can be embedded with distortion 1 into $\mathbb{R}^m$ equipped with the ℓ_1 norm. An ℓ_2^2 distance on X is a distance function for which there there are vectors $\mathbf{v}_x \in \mathbb{R}^m$ for every $x \in X$ so that $d(x,y) = \|\mathbf{v}_x - \mathbf{v}_y\|^2$. If, in addition, d satisfies triangle inequality, we say that d is an ℓ_2^2 metric or negative type metric. It is well known [9] that every ℓ_1 embeddable metric is also a negative type metric.

3. ℓ_1 and integrality gap of SDPs for vertex cover – an "all or nothing" phenomenon. It is well known that for Sparsest Cut there is a tight connection between ℓ_1 embeddability and integrality gap. In fact, the integrality gap is bounded above by the least ℓ_1 distortion of the SDP solution. At the other extreme stand problems like Max Cut and Multi Cut, where ℓ_1 embeddability does not provide any strong evidence for small integrality gap. In this section we show that Vertex Cover falls somewhere between these two classes of ℓ_1-integrality gap relationship witnessing a sharp transition in integrality gap in the following sense: while ℓ_1 embeddability implies no integrality gap, allowing a small distortion, say 1.001, does not prevent an integrality gap of $2 - o(1)$!

Theorem 2. *For a graph $G = (V,E)$, the answer to the SDP formulated in SDP (5) is the size of the minimum vertex cover of G.*

Proof. Let d be the metric solution of SDP (5). We know that d is the result of an ℓ_2^2 unit representation (i.e., it comes from square norms between unit vectors), and furthermore it is ℓ_1 embeddable. By cut representations of ℓ_1 embeddable metrics (see, e.g., [9]) we can assume that there exist $\lambda_t > 0$ and $f_t : \{0\} \cup V \to \{-1, 1\}$, $t = 1, \ldots, m$, such that

$$\|\mathbf{v}_i - \mathbf{v}_j\|^2 = \sum_{t=1}^{m} \lambda_t |f_t(i) - f_t(j)|, \tag{6}$$

for every $i, j \in \{0\} \cup V$. Without loss of generality, we can assume that $f_t(0) = 1$ for every t. For convenience, we switch to talk about Independent Set and its relaxation, which is the same as SDP (5) except the objective becomes Max $\sum_{i \in V} \|\mathbf{v}_0 - \mathbf{v}_i\|^2/4$. Obviously, the theorem follows from showing that this is an exact relaxation.

We argue that (i) $I_t = \{i \in V : f_t(i) = -1\}$ is a (nonempty) independent set for every t, and (ii) $\sum \lambda_t = 2$. Assuming these two statements we get

$$\sum_{i \in V} \frac{\|\mathbf{v}_i - \mathbf{v}_0\|^2}{4} = \sum_{i \in V} \frac{\sum_{t=1}^{m} \lambda_t |1 - f_t(i)|}{4} = \sum_{t=1}^{m} \frac{\lambda_t |I_t|}{2} \leq \max_{t \in [m]} |I_t| \leq \alpha(G),$$

and so the relaxation is exact and we are done.

We now prove the two statements. The first is rather straightforward: For $i, j \in I_t$, (6) implies that $d(i,0) + d(0,j) > d(i,j)$. It follows that ij cannot be an edge or it would violate the first condition of the SDP (we may assume that I_t is nonempty since otherwise the $f_t(\cdot)$ terms have no contribution in (6)). The second statement is more surprising and uses the fact that the solution is optimal. The falsity of such a statement for the problem of MAX CUT explains the different behavior of the latter problem with respect to integrality gaps of ℓ_1 embeddable solutions. We now describe the proof.

Let $\mathbf{v}'_i = (\sqrt{\lambda_1/2} f_1(i), \ldots, \sqrt{\lambda_m/2} f_m(i), 0)$. From (6) we conclude that $\|\mathbf{v}'_i - \mathbf{v}'_j\|^2 = \|\mathbf{v}_i - \mathbf{v}_j\|^2$; hence there exists a vector $\mathbf{w} = (w_1, w_2, \ldots, w_{m+1}) \in \mathbb{R}^{m+1}$ and a linear isometry T on $\operatorname{span}\{\mathbf{v}'_i + \mathbf{w} : 0 \le i \le n\}$ such that

$$\mathbf{v}_i = T\left(\mathbf{v}'_i + \mathbf{w}\right).$$

Since the constraints and the objective function of the SDP are invariant under linear isometries, without loss of generality we may assume that

$$\mathbf{v}_i = \mathbf{v}'_i + \mathbf{w},$$

for $i \in V \cup \{0\}$. We know that

$$1 = \|\mathbf{v}_i\|^2 = \|\mathbf{v}'_i + \mathbf{w}\|^2 = w_{m+1}^2 + \sum_{t=1}^{m} \left(\sqrt{\lambda_t/2} f_t(i) + w_t\right)^2. \tag{7}$$

Since $\|\mathbf{v}'_i\|^2 = \|\mathbf{v}'_0\|^2 = \sum_{t+1}^{m} \lambda_t/2$, for every $i \in V \cup \{0\}$, from (7) we get $\mathbf{v}'_0 \cdot \mathbf{w} = \mathbf{v}'_i \cdot \mathbf{w}$. Summing this over all $i \in V$, we have

$$|V|(\mathbf{v}'_0 \cdot \mathbf{w}) = \sum_{i \in V} \mathbf{v}'_i \cdot \mathbf{w} = \sum_{t=1}^{m} (|V| - 2|I_t|)\sqrt{\lambda_t/2} w_t,$$

or

$$\sum_{t=1}^{m} |V| \sqrt{\lambda_t/2} w_t = \sum_{t=1}^{m} (|V| - 2|I_t|)\sqrt{\lambda_t/2} w_t,$$

and therefore

$$\sum_{t=1}^{m} |I_t| \sqrt{\lambda_t/2} w_t = 0. \tag{8}$$

Now (7) and (8) imply that

$$\max_{t \in [m]} |I_t| \ge \sum_{t=1}^{m} (\sqrt{\lambda_t/2} f_t(0) + w_t)^2 |I_t| = \sum_{t=1}^{m} \left(\frac{\lambda_t |I_t|}{2} + w_t^2 |I_t|\right) \ge \sum_{t=1}^{m} \frac{\lambda_t |I_t|}{2}. \tag{9}$$

As we have observed before

$$\sum_{t=1}^{m} \frac{\lambda_t |I_t|}{2} = \sum_{i \in V} \frac{\|\mathbf{v}_i - \mathbf{v}_0\|^2}{4},$$

which means (as clearly $\sum_{i \in V} \frac{\|\mathbf{v}_i - \mathbf{v}_0\|^2}{4} \ge \alpha(G)$) that the inequalities in (9) must be tight. This implies $w_t^2 |I_t| = 0$, for every $1 \le t \le m$. But since $|I_t| \ne 0$, we get that

$w_t = 0$ for $1 \le t \le m$. Furthermore by (7) and tightness of the first inequality in (9), we get that $w_{m+1} = 0$. Hence $\mathbf{w} = \mathbf{0}$, and then from (7) we get the second statement, i.e., $\sum \lambda_t = 2$. This concludes the proof. □

Now let δ be an arbitrary positive number, and let us relax the last constraint in SDP (5) to get

$$\begin{array}{lll} \textbf{Min} & \sum_{i\in V} 1 - \|\mathbf{v}_0 - \mathbf{v}_i\|^2/4 & \\ \text{s.t.} & \|\mathbf{v}_i - \mathbf{v}_0\|^2 + \|\mathbf{v}_j - \mathbf{v}_0\|^2 = \|\mathbf{v}_i - \mathbf{v}_j\|^2 & \forall\, ij \in E, \\ & \|\mathbf{v}_i\| = 1 & \forall\, i \in \{0\} \cup V, \\ & c_1(\{\mathbf{v}_i : i \in \{0\} \cup V\}, \|\cdot\|^2) \le 1 + \delta. & \end{array}$$

THEOREM 3. *For every $\epsilon > 0$, there is a graph G for which $\frac{\mathrm{vc}(G)}{\mathrm{sd}(G)} \ge 2 - \epsilon$, where $\mathrm{sd}(G)$ is the solution to the above SDP.*

The proof appears in the next section after we describe Charikar's construction.

4. Integrality gap for stronger semidefinite formulations. In this section we discuss the integrality gap for stronger semidefinite formulations of vertex cover. In particular we show that Charikar's construction satisfies both SDPs (11) and (4). We start by describing this construction.

4.1. Charikar's construction. The graphs used in the construction are the so-called Hamming graphs. These are graphs with vertices $\{-1, 1\}^n$, and two vertices are adjacent if their Hamming distance is exactly an even integer $d = \gamma n$. A result of Frankl and Rödl [12] shows that $\mathrm{vc}(G) \ge 2^n - (2 - \delta)^n$, where $\delta > 0$ is a constant depending only on γ. In fact, when one considers the exact dependency of δ in γ it can be shown (see [13]) that as long as $\gamma = \Omega(\sqrt{\log n/n})$, any vertex cover comprises $1 - O(1/n)$ fraction of the graph. Kleinberg and Goemans [22] showed that by choosing a constant γ and n sufficiently large, this graph gives an integrality gap of $2 - \epsilon$ for SDP (1). Charikar [6] showed that in fact G implies the same result for the SDP formulation in (2) too. To this end he introduced the following solution to SDP (2):

For every $\mathbf{u}_i \in \{-1, 1\}^n$, define $\mathbf{u}'_i = \mathbf{u}_i/\sqrt{n}$, so that $\mathbf{u}'_i \cdot \mathbf{u}'_i = 1$. Let $\lambda = 1 - 2\gamma$, $q(x) = x^{2t} + 2t\lambda^{2t-1}x$, and define $\mathbf{y}_0 = (0, \ldots, 0, 1)$, and

$$\mathbf{y}_i = \sqrt{\frac{1-\beta^2}{q(1)}} \left(\underbrace{\mathbf{u}'_i \otimes \ldots \otimes \mathbf{u}'_i}_{2t \text{ times}}, \sqrt{2t\lambda^{2t-1}}\mathbf{u}'_i, 0 \right) + \beta \mathbf{y}_0,$$

where β will be determined later. Note that $\mathbf{y}_i$ is normalized to satisfy $\|\mathbf{y}_i\| = 1$.

Moreover $\mathbf{y}_i$ is defined so that $\mathbf{y}_i \cdot \mathbf{y}_j$ takes its minimum value when $ij \in E$, i.e., when $\mathbf{u}'_i \cdot \mathbf{u}'_j = -\lambda$. As is shown in [6], for every $\epsilon > 0$ we may set $t = \Omega(\frac{1}{\epsilon}), \beta = \Theta(1/t), \gamma = \frac{1}{4t}$ to get that $(\mathbf{y}_0 - \mathbf{y}_i) \cdot (\mathbf{y}_0 - \mathbf{y}_j) = 0$ for $ij \in E$, while $(\mathbf{y}_0 - \mathbf{y}_i) \cdot (\mathbf{y}_0 - \mathbf{y}_j) \ge 0$ always.

Now we verify that all the triangle inequalities; i.e., the second constraint of SDP (2) is satisfied: First note that since every coordinate takes only two different values for the vectors in $\{\mathbf{y}_i : i \in V\}$, it is easy to see that $c_1(\{\mathbf{y}_i : i \in V\}, \|\cdot\|^2) = 1$. So the triangle inequality holds when $i, j, k \in V$. When $i = 0$ or $j = 0$, the inequality is trivial, and it only remains to verify the case that $k = 0$, i.e., $(\mathbf{y}_0 - \mathbf{y}_i) \cdot (\mathbf{y}_0 - \mathbf{y}_j) \ge 0$, which was already mentioned above. Now $\sum_{i\in V}(1 + \mathbf{y}_0 \cdot \mathbf{y}_i)/2 = \frac{1+\beta}{2} \cdot |V| = \left(\frac{1}{2} + O(\epsilon)\right)|V|$. In our application, we prefer to set γ and ϵ to be $\Omega(\sqrt{\frac{\log\log n}{\log n}})$ and

since, by the above comment, $\mathrm{vc}(G) = (1 - O(1/n))|V|$ the integrality gap we get is

$$(1 - O(1/n))/(1/2 + O(\epsilon)) = 2 - O(\epsilon) = 2 - O\left(\sqrt{\frac{\log\log|V|}{\log|V|}}\right).$$

4.2. Proof of Theorem 3. We show that the negative type metric implied by Charikar's solution (after adjusting the parameters appropriately) requires distortion of at most $1 + \delta$. Let $\mathbf{y}_i$ and $\mathbf{u}'_i$ be defined as in section 4.1. To prove Theorem 3, it is sufficient to prove that $c_1(\{\mathbf{y}_i : i \in \{0\} \cup V\}, \|\cdot\|^2) = 1 + o(1)$. Note that every coordinate of $\mathbf{y}_i$ for all $i \in V$ takes at most two different values. It is easy to see that this implies $c_1(\{\mathbf{y}_i : i \in V\}, \|\cdot\|^2) = 1$. In fact,

$$f : \mathbf{y}_i \mapsto \frac{1-\beta^2}{q(1)}\left(\frac{2}{n^t}\underbrace{\mathbf{u}'_i \otimes \ldots \otimes \mathbf{u}'_i}_{2t \text{ times}}, \frac{2}{\sqrt{n}} 2t\lambda^{2t-1}\mathbf{u}'_i\right) \tag{10}$$

is an isometry from $(\{\mathbf{y}_i : i \in V\}, \|\cdot\|^2)$ to ℓ_1. For $i \in V$, we have

$$\|f(\mathbf{y}_i)\|_1 = \frac{1-\beta^2}{q(1)}\left(\frac{2}{n^t} \times \frac{n^{2t}}{n^t} + \frac{2}{\sqrt{n}} 2t\lambda^{2t-1}\frac{1}{\sqrt{n}} + 0\right) = \frac{1-\beta^2}{q(1)} \times \left(2 + 4t\lambda^{2t-1}\right).$$

Since $\beta = \Theta(\frac{1}{t})$, recalling that $\lambda = 1 - \frac{1}{2t}$, it is easy to see that for every $i \in V$, $\lim_{t\to\infty} \|f(\mathbf{y}_i)\|_1 = 2$. On the other hand, for every $i \in V$

$$\lim_{t\to\infty} \|\mathbf{y}_i - \mathbf{y}_0\|^2 = \lim_{t\to\infty} 2 - 2(\mathbf{y}_i \cdot \mathbf{y}_0) = \lim_{t\to\infty} 2 - 2\beta = 2.$$

So if we extend f to $\{\mathbf{y}_i : i \in V \cup \{0\}\}$ by defining $f(\mathbf{y}_0) = \mathbf{0}$, we obtain a mapping from $(\{\mathbf{y}_i : i \in V \cup \{0\}\}, \|\cdot\|^2)$ to ℓ_1 whose distortion tends to 1 as t goes to infinity.

4.3. Karakostas' and pentagonal SDP formulations. Karakostas suggests the following SDP relaxation, which is the result of adding to SDP (3) the triangle inequalities applied to the set $\{\mathbf{v}_i : i \in V \cup \{0\}\} \cup \{-\mathbf{v}_i : i \in V \cup \{0\}\}$.

$$\begin{array}{lll}
\textbf{Min} & \sum_{i\in V}(1 + \mathbf{v}_0\mathbf{v}_i)/2 & \\
\text{s.t.} & (\mathbf{v}_i - \mathbf{v}_0)\cdot(\mathbf{v}_j - \mathbf{v}_0) = 0 & \forall\ ij \in E \\
& (\mathbf{v}_i - \mathbf{v}_k)\cdot(\mathbf{v}_j - \mathbf{v}_k) \geq 0 & \forall\ i,j,k \in V \\
& (\mathbf{v}_i + \mathbf{v}_k)\cdot(\mathbf{v}_j - \mathbf{v}_k) \geq 0 & \forall\ i,j,k \in V \\
& (\mathbf{v}_i + \mathbf{v}_k)\cdot(\mathbf{v}_j + \mathbf{v}_k) \geq 0 & \forall\ i,j,k \in V \\
& \|\mathbf{v}_i\| = 1 & \forall\ i \in \{0\} \cup V.
\end{array} \tag{11}$$

THEOREM 4. *The integrality gap of SDP* (11) *is* $2 - O(\sqrt{\log\log|V|/\log|V|})$.

Proof. We show that Charikar's construction satisfies formulation (11). By [6] and from the discussion in section 4.1, it follows that all edge constraints and triangle inequalities of the original points hold. Hence we need only consider triangle inequalities with at least one nonoriginal point. By homogeneity, we may assume that there is exactly one such point.

Since all coordinates of $\mathbf{y}_i$ for $i > 0$ assume only two values with the same absolute value, it is clear that not only is the metric they induce ℓ_1, but also taking $\pm\mathbf{y}_i$ for $i > 0$ gives an ℓ_1 metric; in particular, all triangle inequalities that involve these vectors are satisfied. In fact, we may fix our attention to triangles in which $\pm\mathbf{y}_0$ is the middle point. This is since

$$(\pm\mathbf{y}_i - \pm\mathbf{y}_j)\cdot(\mathbf{y}_0 - \pm\mathbf{y}_j) = (\pm\mathbf{y}_j - \mathbf{y}_0)\cdot(\mp\mathbf{y}_i - \mathbf{y}_0).$$

Consequently, and using symmetry, we are left with checking the nonnegativity of $(\mathbf{y}_i + \mathbf{y}_0) \cdot (\mathbf{y}_j + \mathbf{y}_0)$ and $(-\mathbf{y}_i - \mathbf{y}_0) \cdot (\mathbf{y}_j - \mathbf{y}_0)$.

$$(\mathbf{y}_i+\mathbf{y}_0)\cdot(\mathbf{y}_j+\mathbf{y}_0) = 1+\mathbf{y}_0\cdot(\mathbf{y}_i+\mathbf{y}_j)+\mathbf{y}_i\cdot\mathbf{y}_j \geq 1+2\beta+\beta^2-\left(1-\beta^2\right) = 2\beta(1+\beta) \geq 0.$$

Finally, $(-\mathbf{y}_i - \mathbf{y}_0) \cdot (\mathbf{y}_j - \mathbf{y}_0) = 1 + \mathbf{y}_0 \cdot (\mathbf{y}_i - \mathbf{y}_j) - \mathbf{y}_i \cdot \mathbf{y}_j = 1 - \mathbf{y}_i \cdot \mathbf{y}_j \geq 0$ as $\mathbf{y}_i, \mathbf{y}_j$ are of norm 1. □

By now we know that taking all the ℓ_1 constraints leads to an exact relaxation, but not a tractable one. Our goal here is to explore the possibility that stepping towards ℓ_1 embeddability while still maintaining computational feasibility would considerably reduce the integrality gap. A canonical subset of valid inequalities for ℓ_1 metrics is the so-called *Hypermetric inequalities*. Metrics that satisfy all these inequalities are called *hypermetrics*. Again, taking all these constraints is not feasible, and yet we do not know whether this may lead to a better integrality gap (notice that we do not know that Theorem 2 remains true if we replace the ℓ_1 embeddability constraints with a hypermetricity constraint). See [9] for a related discussion about hypermetrics. We instead consider the effect of adding a small number of such constraints. The simplest hypermetric inequalities beside triangle inequalities are the *pentagonal* inequalities. These constraints consider two sets of points of size 2 and 3, and require that the sum of the distances between points in different sets is at least the sum of the distances within sets. Formally, let $S, T \subset X$, $|S| = 2, |T| = 3$, and then we have the inequality $\sum_{i\in S, j\in T} d(i,j) \geq \sum_{i,j\in S} d(i,j) + \sum_{i,j\in T} d(i,j)$. To appreciate this inequality it is useful to describe where it fails. Consider the graph metric of $K_{2,3}$. Here, the LHS of the inequality is 6 and the RHS is 8; hence $K_{2,3}$ violates the pentagonal inequality. In the following theorem we show that this strengthening past the triangle inequalities fails to reduce the integrality gap significantly.

THEOREM 5. *The integrality gap of SDP* (4) *is* $2 - O(\sqrt{\log\log|V|/\log|V|})$.

Proof. We note that in order to satisfy the triangle inequalities, the conditions that should be satisfied by the "tensoring-polynomial" used in the construction ("q" in the notation of the previous subsection) are rather modest. Essentially we needed that $q'(-\lambda) = 0$, $q(-\lambda)/q(1)$ approaches -1, and that $q''(-\lambda) \geq 0$. For the pentagonal inequalities we need to require more properties from q, namely that it is convex on its entire domain and that its derivative satisfies certain linear conditions, all of which turn out to be true.

We show that the metric space used in Charikar's construction is a feasible solution. By ignoring $\mathbf{y}_0$ the space defined by $d(i,j) = \|\mathbf{y}_i - \mathbf{y}_j\|^2$ is ℓ_1 embeddable. Therefore, the only ℓ_1-valid inequalities that may be violated are ones containing $\mathbf{y}_0$. Hence, we wish to consider a pentagonal inequality containing $\mathbf{y}_0$ and four other vectors, denoted by $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4$. Assume first that the partition of the five points in the inequality puts $\mathbf{y}_0$ together with two other points; then, using the fact that $d(0,1) = d(0,2) = d(0,3) = d(0,4)$ and triangle inequality we get that such an inequality must hold. It remains to consider a partition of the form $(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}, \{\mathbf{y}_4, \mathbf{y}_0\})$, and show that

$$d(1,2)+d(1,3)+d(2,3)+d(0,4) \leq d(1,4)+d(2,4)+d(3,4)+d(0,1)+d(0,2)+d(0,3).$$

As the vectors are of unit norm, it is clear that $d(0,i) = 2 - 2\beta$ for all $i > 0$ and that $d(i,j) = 2 - 2\mathbf{y}_i\mathbf{y}_j$. Recall that every $\mathbf{y}_i$ is associated with a $\{-1,1\}$ vector $\mathbf{u}_i$ and with its normalized multiple $\mathbf{u}'_i$. Also, it is simple to check that $\mathbf{y}_i \cdot \mathbf{y}_j = \beta^2 + (1-\beta^2)q(\mathbf{u}'_i \cdot \mathbf{u}'_j)/q(1)$ where $q(x) = x^{2t} + 2\lambda^{2t-1}x$. After substituting the

distances as functions of the normalized vectors, our goal will then be to show

$$E = q(\mathbf{u}_1' \cdot \mathbf{u}_2') + q(\mathbf{u}_1' \cdot \mathbf{u}_3') + q(\mathbf{u}_2' \cdot \mathbf{u}_3') - q(\mathbf{u}_1' \cdot \mathbf{u}_4') - q(\mathbf{u}_2' \cdot \mathbf{u}_4') - q(\mathbf{u}_3' \cdot \mathbf{u}_4') \geq -\frac{2q(1)}{1+\beta}. \tag{12}$$

The rest of the proof analyzes the minima of the function E and ensures that (12) is satisfied at those minima. We first partition the coordinates of the original hypercube into four sets according to the values assumed by $\mathbf{u}_1, \mathbf{u}_2$, and $\mathbf{u}_3$. We may assume that in any coordinate at most one of these get the value 1 (otherwise multiply the values of the coordinate by -1). We get four sets, P_0 for the coordinates in which all three vectors assume value -1, and P_1, P_2, P_3 for the coordinates in which exactly $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$, respectively, assumes value 1.

We now consider $\mathbf{u}_4$. We argue that without loss of generality, we may assume that $\mathbf{u}_4$ is "pure" on each of the P_0, P_1, P_2, P_3 at a minimum of E; in other words it is either all 1 or all -1 on each one of P_0, P_1, P_2, P_3.

PROPOSITION 1. *If there is a violating configuration, then there is one in which* $\mathbf{u}_4$ *is either all* 1 *or all* -1 *on each one of* P_0, P_1, P_2, P_3.

Proof. Assume for the sake of contradiction that there are w coordinates in P_0 on which $\mathbf{u}_4$ assumes value -1, and that $0 < w < |P_0|$. Let $\mathbf{u}_4^+$ (similarly $\mathbf{u}_4^-$) be identical to $\mathbf{u}_4$ except we replace one 1 in P_0 by -1 (replace one -1 in P_0 by 1). We show that replacing $\mathbf{u}_4$ by $\mathbf{u}_4^+$ or by $\mathbf{u}_4^-$ we decrease the expression E. Let $p_i = \mathbf{u}_i \cdot \mathbf{u}_4$, $p_i^+ = \mathbf{u}_i' \cdot (\mathbf{u}_4^+)'$, and $p_i^- = \mathbf{u}_i' \cdot (\mathbf{u}_4^-)'$ for $i = 1, 2, 3$. Notice that the above replacement changes only the negative terms in (12) so our goal now is to show that $\sum_{i=1}^3 q(p_i) < \max\{\sum_{i=1}^3 q(p_i^+), \sum_{i=1}^3 q(p_i^-)\}$. But

$$\max\left\{\sum_{i=1}^3 q(p_i^+), \sum_{i=1}^3 q(p_i^-)\right\} \geq \sum_{i=1}^3 \frac{q(p_i^+) + q(p_i^-)}{2} > \sum_{i=1}^3 q\left(\frac{p_i^+ + p_i^-}{2}\right) = \sum_{i=1}^3 q(p_i),$$

where the last inequality is using the (strict) convexity of q. This of course applies to P_1, P_2, and P_3 in precisely the same manner. ∎

For P_0, we can in fact say something stronger than we do for P_1, P_2, P_3:

PROPOSITION 2. *If there is a violating configuration, then there is one in which* $\mathbf{u}_4$ *has all the* P_0 *coordinates set to* -1.

The above characterizations significantly limit the type of configurations we need to check. Proposition 1 was based solely on the (strict) convexity of q. Proposition 2 is more involved and uses more properties of the polynomial q. If q was a monotone increasing function it would be obvious, but of course the whole point behind q is that it brings to minimum some intermediate value $(-\lambda)$ and hence cannot be increasing. We postpone the proof of Proposition 2 till the end of the section, and we will now continue our analysis assuming the proposition.

The cases that are left are characterized by whether $\mathbf{u}_4$ is 1 or -1 on each of P_1, P_2, P_3. By symmetry all we really need to know is $\xi(\mathbf{u}_4) = |\{i \ : \ \mathbf{u}_4 \text{ is 1 on } P_i\}|$. If $\xi(\mathbf{u}_4) = 1$ it means that $\mathbf{u}_4$ is the same as one of $\mathbf{u}_1, \mathbf{u}_2$, or $\mathbf{u}_4$; hence the pentagonal inequality reduces to the triangle inequality, which we already know is valid. If $\xi(\mathbf{u}_4) = 3$, it is easy to see that $\mathbf{u}_1'\mathbf{u}_4' = \mathbf{u}_2'\mathbf{u}_3'$, and likewise $\mathbf{u}_2'\mathbf{u}_4' = \mathbf{u}_1'\mathbf{u}_3'$ and $\mathbf{u}_3'\mathbf{u}_4' = \mathbf{u}_1'\mathbf{u}_2'$; hence E is 0 for these cases, which means that (12) is satisfied.

We are left with the cases $\xi(\mathbf{u}_4) \in \{0, 2\}$.

Case 1: $\xi(\mathbf{u}_4) = 0$. Let $x = \frac{2}{n}|P_1|, y = \frac{2}{n}|P_2|$, and $z = \frac{2}{n}|P_3|$. Notice that $x + y + z = \frac{2}{n}(|P_1| + |P_2| + |P_3|) \leq 2$, as these sets are disjointed. Now, think of

$$E = q(1-(x+y)) + q(1-(x+z)) + q(1-(y+z)) - q(1-x) - q(1-y) - q(1-z)$$

as a function from $\mathbb{R}^3$ to $\mathbb{R}$. We will show that E achieves its minimum at points where either x, y or z are zero. Assume that $0 \le x \le y \le z$.

Consider the function $g(\delta) = E(x-\delta, y+\delta, z)$. It is easy to see that $g'(0) = q'(1-(x+z)) - q'(1-(y+z)) - q'(1-x) + q'(1-y)$. We will prove that $g'(\delta) \le 0$ for every $\delta \in [0, x]$. This, by the mean value theorem, implies that $E(0, x+y, z) \le E(x, y, z)$, and hence we may assume that $x = 0$. This means that $\mathbf{y}_1 = \mathbf{y}_4$ which reduces to the triangle inequality on $\mathbf{y}_0, \mathbf{y}_2, \mathbf{y}_3$.

Note that in $g'(0)$, the two arguments in the terms with positive signs have the same average as the arguments in the terms with negative signs, namely, $\mu = 1 - (x+y+z)/2$. We now have $g'(0) = q'(\mu+b) - q'(\mu+s) - q'(\mu-s) + q'(\mu-b)$, where $b = (x-y+z)/2, s = (-x+y+z)/2$. After calculations:

$$\begin{aligned} g'(0) &= 2t\left[(\mu+b)^{2t-1} + (\mu-b)^{2t-1} - (\mu+s)^{2t-1} - (\mu-s)^{2t-1}\right] \\ &= 4t \sum_{i \text{ even}} \binom{2t-1}{i} \mu^{2t-1-i}(b^i - s^i). \end{aligned}$$

Observe that $\mu \ge 0$. Since $x \le y$, we get that $s \ge b \ge 0$. This means that $g'(0) \le 0$. It can be easily checked that the same argument holds if we replace x, y by $x-\delta$ and $y+\delta$. Hence $g'(\delta) \le 0$ for every $\delta \in [0, x]$, and we are done.

Case 2: $\xi(\mathbf{u}_4) = 2$. The expression for E is now

$$\begin{aligned} E = {} & q(1-(x+y)) + q(1-(x+z)) + q(1-(y+z)) - q(1-x) \\ & - q(1-y) - q(1-(x+y+z)) \end{aligned}$$

Although $E(x, y, z)$ is different than in Case 1, the important observation is that if we consider again the function $g(\delta) = E(x-\delta, y+\delta, z)$, then the derivative $g'(\delta)$ is the same as in Case 1, and hence the same analysis shows that $E(0, x+y, z) \le E(x, y, z)$. But if $x = 0$, then $\mathbf{y}_2$ identifies with $\mathbf{y}_4$ and the inequality reduces to the triangle inequality on $\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_3$. ∎

To complete the proof, it remains to prove Proposition 2.

Proof of Proposition 2. Fix a configuration for $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ and as before let $x = \frac{2}{n}|P_1|$, $y = \frac{2}{n}|P_2|$, $z = \frac{2}{n}|P_3|$, and $w = \frac{2}{n}|P_0|$, where $w > 0$. Consider a vector $\mathbf{u}_4$ that has all -1's in P_0. Let $H_i = \frac{2}{n}H(\mathbf{u}_i, \mathbf{u}_4)$, where $H(\mathbf{u}_i, \mathbf{u}_4)$ is the Hamming distance from $\mathbf{u}_4$ to $\mathbf{u}_i$, $i = 1, 2, 3$. It suffices to show that replacing the P_0 part of $\mathbf{u}_4$ with 1's (which means adding w to each H_i) does not decrease the LHS of (12), i.e.,

$$\begin{aligned} q(1-H_1) + q(1-H_2) + q(1-H_3) \ge {} & q(1-(H_1+w)) + q(1-(H_2+w)) \\ & + q(1-(H_3+w)). \end{aligned} \tag{13}$$

Because of the convexity of q, the cases that we need to consider are characterized by whether $\mathbf{u}_4$ is 1 or -1 on each of P_1, P_2, P_3. By symmetry there are four cases to check, corresponding to the different values of $\xi(\mathbf{u}_4)$. In most of these cases, we use the following argument: consider the function $g(\delta) = q(1-(H_1+\delta)) + q(1-(H_2+\delta)) + q(1-(H_3+\delta))$, where $\delta \in [0, w]$. Let $a_i = 1-(H_i+\delta)$. The derivative $g'(\delta)$ is

$$g'(\delta) = -(q'(a_1) + q'(a_2) + q'(a_3)) = -2t\left(a_1^{2t-1} + a_2^{2t-1} + a_3^{2t-1} + 3\lambda^{2t-1}\right).$$

If we show that the derivative is negative for any $\delta \in [0, w]$, that would imply that $g(0) \ge g(w)$ and hence we are done since we have a more violating configuration if we do not add w to the Hamming distances.

Case 1: $\xi(\mathbf{u}_4) = 0$. In this case $H_1 = x$, $H_2 = y$, and $H_3 = z$. Note that $x + y + z + w = 2$. Hence, if $H_i \ge 1$ for some i, say for H_1, then $H_2 + \delta \le 1$ and $H_3 + \delta \le 1$.

This implies that $a_2 \geq 0$ and $a_3 \geq 0$. Thus

$$g'(\delta) \leq -\left(-1 + 3\lambda^{2t-1}\right) \leq 1 - 3/e < 0$$

since $\lambda^{2t-1} = (1 - \frac{1}{2t})^{2t-1} \geq 1/e$. Hence we are done.

Therefore, we can assume that $H_i < 1$ for all i, i.e., $1 - H_i \geq 0$. We now compare the LHS and RHS of (13). In particular we claim that each term $q(1 - H_i)$ is at least as big as the corresponding term $q(1 - (H_i + w))$. This is because of the form of the function q. Note that q is increasing in $[0, 1]$ and also that the value of q at any point $x \in [0, 1]$ is greater than the value of q at any point $y \in [-1, 0)$. Therefore since $1 - H_i > 0$ and since we only subtract w from each point, it follows that (13) holds.

Case 2: $\xi(\mathbf{u}_4) = 1$. Assume without loss of generality that $\mathbf{u}_4$ is 1 on P_1 only. In this case, $H_1 = 0$, $H_2 = x + y$, and $H_3 = x + z$. The LHS of inequality (13) is now

$$LHS = q(1) + q(1 - (x + y)) + q(1 - (x + z)),$$

whereas the RHS is

$$\begin{aligned} RHS &= q(1 - w) + q(1 - (x + y + w)) + q(1 - (x + z + w)) \\ &= q(1 - w) + q(-1 + z) + q(-1 + y) \end{aligned}$$

by using the fact that $x + y + w = 2 - z$.

Let $\alpha_1 = 1$, $\alpha_2 = 1-(x+y)$, and $\alpha_3 = 1-(x+z)$. The LHS is the sum of the values of q at these points whereas the RHS is the sum of the values of q after shifting each point α_i to the left by w. Let $\alpha_i' = \alpha_i - w$. The difference $\Delta = q(1) - q(1 - w)$ will always be positive since $q(1)$ is the highest value that q achieves in $[-1, 1]$. Therefore, to show that (13) holds it is enough to show that the potential gain in q from shifting α_2 and α_3 is at most Δ. Suppose not and consider such a configuration. This means that either $q(\alpha_2') > q(\alpha_2)$ or $q(\alpha_3') > q(\alpha_3)$ or both. We consider the case that both points achieve a higher value after being shifted. The same arguments apply if we have only one point that improves its value. Hence we assume that $q(\alpha_2') > q(\alpha_2)$ and $q(\alpha_3') > q(\alpha_3)$. Before we proceed, we state some properties of q, which can be easily verified.

Claim 1. *The function q is decreasing in $[-1, -\lambda]$ and increasing in $[-\lambda, 1]$. Furthermore, for any 2 points x, y such that $x \in [-1, 2 - 3\lambda]$ and $y \geq 2 - 3\lambda$, $q(y) \geq q(x)$.*

Using the above claim, we argue about the location of α_2 and α_3. If $\alpha_2 \geq 2 - 3\lambda \geq -\lambda$, then $q(\alpha_2) \geq q(\alpha_2')$. Thus both α_2 and α_3 must belong to $[-1, 2-3\lambda] = [-1, -1 + \frac{3}{2t}]$. We will restrict further the location of α_2 and α_3 by making some more observations about q. The interval $[-1, 2 - 3\lambda]$ is the union of $A_1 = [-1, -\lambda]$ and $A_2 = [-\lambda, 2 - 3\lambda]$, and we know q is decreasing in A_1 and increasing in A_2. We claim that α_2, α_3 should belong to A_1 in the worst possible violation of (13). To see this, suppose $\alpha_2 \in A_2$ and $\alpha_3 \in A_2$ (the case with $\alpha_2 \in A_2$, $\alpha_3 \in A_1$ can be handled similarly). We know that q is the sum of a linear function and the function x^{2t}. Hence when we shift the 3 points to the left, the difference $q(1) - q(1 - w)$ is at least as big as a positive term that is linear in w. This difference has to be counterbalanced by the differences $q(\alpha_2') - q(\alpha_2)$ and $q(\alpha_3') - q(\alpha_3)$. However, the form of q ensures that there is a point $\zeta_2 \in A_1$ such that $q(\alpha_2) = q(\zeta_2)$ and ditto for α_3. By considering the configuration where $\alpha_2 \equiv \zeta_2$ and $\alpha_3 \equiv \zeta_3$ we will have the same contribution from the terms $q(\alpha_2') - q(\alpha_2)$ and $q(\alpha_3') - q(\alpha_3)$ and at the same time a smaller w.

Therefore, we may assume that $w \leq |A_1| = \frac{1}{2t}$. By substituting the value of q, (13) is equivalent to showing that

$$1 - (1 - w)^{2t} + 6t\lambda^{2t-1}w \geq (\alpha_2 - w)^{2t} - \alpha_2^{2t} + (\alpha_3 - w)^{2t} - \alpha_3^{2t}.$$

It is easy to see that the difference $1-(1-w)^{2t}$ is greater than or equal to the difference $(\alpha_2-w)^{2t}-\alpha_2^{2t}$ by convexity. Hence it suffices to show

$$6t\lambda^{2t-1}w \geq (\alpha_3-w)^{2t}-\alpha_3^{2t}.$$

We know that the LHS is at least $(6t/e)w$. The difference $(\alpha_3-w)^{2t}-\alpha_3^{2t}$ can be estimated using the derivatives of x^{2t} and turns out to be at most $(6t/e)w$. Therefore, no configuration in this case can violate (13).

Case 3: $\xi(\mathbf{u}_4)=2$. Assume that $\mathbf{u}_4$ is 1 on P_1 and P_2. Now $H_1=y$, $H_2=x$, and $H_3=x+y+z$. The LHS and RHS of (13) are

$$LHS = q(1-y)+q(1-x)+q(1-(x+y+z)),$$
$$RHS = q(1-(y+w))+q(1-(x+w))+q(-1).$$

As in case 2, let $\alpha_1=1-y$, $\alpha_2=1-x$, and $\alpha_3=1-(x+y+z)$ be the three points before shifting by w. First note that either $\alpha_1>0$ or $\alpha_2>0$. This comes from the constraint that $x+y+z+w=2$. Assume that $\alpha_1>0$. Hence $q(\alpha_1)-q(\alpha_1-w)>0$. If $\alpha_2\notin[-1,2-3\lambda]$, then we would be done because by the above claim, $q(\alpha_2)-q(\alpha_2-w)>0$. Therefore, the only way that (13) can be violated is if the nonlinear term $(\alpha_3-w)^{2t}-\alpha_3^{2t}$ can compensate for the loss for the other terms. It can be easily checked that this cannot happen. Hence we may assume that both $\alpha_2,\alpha_3\in[-1,2-3\lambda]$ and that $q(\alpha_2-w)>q(\alpha_2)$, $q(\alpha_3-w)>q(\alpha_3)$. The rest of the analysis is based on arguments similar to case 2 and we omit it.

Case 4: $\xi(\mathbf{u}_4)=3$. This case can be handled using similar arguments to cases 2 and 3.

5. Lower bound for embedding negative type metrics into ℓ_1. While, in view of Theorem 3, Charikar's metric does not supply an example that is far from ℓ_1, we may still (partly motivated by Theorem 2) utilize the idea of "tensoring the cube" and then adding some more points in order to achieve negative type metrics that are not ℓ_1 embeddable. Our starting point is an isoperimetric inequality on the cube that generalizes the standard one. Such a setting is also relevant in [21, 23] where harmonic analysis tools are used to bound expansion; these tools are unlikely to be applicable to our case where the interest and improvements lie in the constants.

THEOREM 1 (generalized isoperimetric inequality). *For every set $S\subseteq Q_n$,*

$$|E(S,S^c)| \geq |S|(n-\log_2|S|)+p(S),$$

where $p(S)$ denotes the number of vertices $\mathbf{u}\in S$ such that $-\mathbf{u}\in S$.

Proof. We use induction on n. Divide Q_n into two sets $V_1=\{\mathbf{u}:\mathbf{u}_1=1\}$ and $V_{-1}=\{\mathbf{u}:\mathbf{u}_1=-1\}$. Let $S_1=S\cap V_1$ and $S_{-1}=S\cap V_{-1}$. Now, $E(S,S^c)$ is the disjoint union of $E(S_1,V_1\setminus S_1)$, $E(S_{-1},V_{-1}\setminus S_{-1})$, and $E(S_1,V_{-1}\setminus S_{-1})\cup E(S_{-1},V_1\setminus S_1)$. Define the operator $\widehat{\cdot}$ on Q_n to be the projection onto the last $n-1$ coordinates, so, for example, $\widehat{S_1}=\{\mathbf{u}\in Q_{n-1}:(1,\mathbf{u})\in S_1\}$. It is easy to observe that $|E(S_1,V_{-1}\setminus S_{-1})\cup E(S_{-1},V_1\setminus S_1)|=|\widehat{S_1}\Delta\widehat{S_{-1}}|$. Without loss of generality assume that $|S_1|\geq|S_{-1}|$. We argue that

$$p(S)+|S_1|-|S_{-1}| \leq p\left(\widehat{S_1}\right)+p\left(\widehat{S_{-1}}\right)+\left|\widehat{S_1}\Delta\widehat{S_{-1}}\right|. \tag{14}$$

To prove (14), for every $\mathbf{u}\in\{-1,1\}^{n-1}$, we show that the contribution of $(1,\mathbf{u})$, $(1,-\mathbf{u})$, $(-1,\mathbf{u})$, and $(-1,-\mathbf{u})$ to the right-hand side of (14) is at least as large as

their contribution to the left-hand side: This is trivial if the contribution of these four vectors to $p(S)$ is not more than their contribution to $p(\widehat{S_1})$, and $p(\widehat{S_{-1}})$. We therefore assume that the contribution of the four vectors to $p(S)$, $p(\widehat{S_1})$, and $p(\widehat{S_{-1}})$ are 2, 0, and 0, respectively. Then without loss of generality we may assume that $(1,\mathbf{u}),(-1,-\mathbf{u}) \in S$ and $(1,-\mathbf{u}),(-1,\mathbf{u}) \notin S$, and in this case the contribution to both sides is 2. By induction hypothesis and (14) we get

$$\begin{aligned}
|E(S,S^c)| &= |E(\widehat{S_1}, Q_{n-1} \setminus \widehat{S_1})| + |E(\widehat{S_{-1}}, Q_{n-1} \setminus \widehat{S_{-1}})| + |\widehat{S_1}\Delta\widehat{S_{-1}}| \\
&\geq |S_1|(n-1-\log_2|S_1|) + p(\widehat{S_1}) + |S_{-1}|(n-1-\log_2|S_{-1}|) \\
&\quad + p(\widehat{S_{-1}}) + |\widehat{S_1}\Delta\widehat{S_{-1}}| \\
&\geq |S|n - |S| - (|S_1|\log_2|S_1| + |S_{-1}|\log_2|S_{-1}|) + p(\widehat{S_1}) \\
&\quad + p(\widehat{S_{-1}}) + |\widehat{S_1}\Delta\widehat{S_{-1}}| \\
&\geq |S|n - (2|S_{-1}| + |S_1|\log_2|S_1| + |S_{-1}|\log_2|S_{-1}|) + p(S).
\end{aligned}$$

Now the lemma follows from the fact that $2|S_{-1}| + |S_1|\log_2|S_1| + |S_{-1}|\log_2|S_{-1}| \leq |S|\log_2|S|$, which can be obtained from the assumption $|S_{-1}| \leq |S_1|$ using easy calculus. ∎

We call a set $S \subseteq Q_n$ *symmetric* if $-\mathbf{u} \in S$ whenever $\mathbf{u} \in S$. Note that $p(S) = |S|$ for symmetric sets S.

COROLLARY 1. *For every symmetric set* $S \subseteq Q_n$

$$|E(S,S^c)| \geq |S|(n - \log_2|S| + 1).$$

The corollary above implies the following Poincaré inequality.

PROPOSITION 3 (Poincaré inequality for the cube and an additional point).

Let $f : Q_n \cup \{\mathbf{0}\} \to \mathbb{R}^m$ *satisfy that* $f(\mathbf{u}) = f(-\mathbf{u})$ *for every* $\mathbf{u} \in Q_n$, *and let* $\alpha = \frac{\ln 2}{14 - 8\ln 2}$.

Then the following Poincaré inequality holds.

$$\frac{1}{2^n}\cdot\frac{4}{7}(4\alpha+1/2)\sum_{\mathbf{u},\mathbf{v}\in Q_n}\|f(\mathbf{u})-f(\mathbf{v})\|_1 \leq \alpha\sum_{\mathbf{uv}\in E}\|f(\mathbf{u})-f(\mathbf{v})\|_1 + \frac{1}{2}\sum_{\mathbf{u}\in Q_n}\|f(\mathbf{u})-f(\mathbf{0})\|_1$$

Proof. It is enough to prove the above inequality for $f : V \to \{0,1\}$. We may assume without loss of generality that $f(\mathbf{0}) = 0$. Associating S with $\{\mathbf{u} : f(\mathbf{u}) = 1\}$, the inequality of the proposition reduces to

$$\frac{1}{2^n}\frac{8}{7}(4\alpha + 1/2)|S||S^c| \leq \alpha|E(S,S^c)| + |S|/2, \tag{15}$$

where S is a symmetric set, owing to the condition $f(\mathbf{u}) = f(-\mathbf{u})$. From the isoperimetric inequality of Theorem 1 we have that $|E(S,S^c)| \geq |S|(x+1)$ for $x = n - \log_2|S|$ and so

$$\left(\frac{\alpha(x+1)+1/2}{1-2^{-x}}\right)\frac{1}{2^n}|S||S^c| \leq \alpha|E(S,S^c)| + |S|/2.$$

Lemma 1 below shows that $\frac{\alpha(x+1)+1/2}{1-2^{-x}}$ attains its minimum in $[1,\infty)$ at $x = 3$ when $\frac{\alpha(x+1)+1/2}{1-2^{-x}} \geq \frac{4\alpha+1/2}{7/8}$, and Inequality (15) is proven. ∎

LEMMA 1. *The function* $f(x) = \frac{\alpha(x+1)+1/2}{1-2^{-x}}$ *for* $\alpha = \frac{\ln 2}{14-8\ln 2}$ *attains its minimum in* $[1,\infty]$ *at* $x = 3$.

Proof. The derivative of f is

$$\frac{\alpha(1-2^{-x})-(\alpha(x+1)+1/2)\ln(2)2^{-x}}{(1-2^{-x})^2}.$$

It is easy to see that $f'(3)=0$, $f(1)=4\alpha+1>8/7$, and $\lim_{x\to\infty} f(x)=\infty$. So it is sufficient to show that

$$g(x)=1-2^{-x}-(\alpha(x+1)+1/2)\ln(2)2^{-x}$$

is an increasing function in the interval $[1,\infty)$. To show this note that

$$g'(x)=2^{-x}\ln(2)\left(1-\alpha+\alpha x\ln(2)+\alpha\ln(2)\right)>0$$

for $x\geq 1$. ∎

THEOREM 6. *Let $V=\{\tilde{\mathbf{u}}:\mathbf{u}\in Q_n\}\cup\{\mathbf{0}\}$, where $\tilde{\mathbf{u}}=\mathbf{u}\otimes\mathbf{u}$. Then for the semi-metric space $X=(V,\|\cdot\|^2)$ we have $c_1(X)\geq\frac{8}{7}-\epsilon$, for every $\epsilon>0$ and sufficiently large n.*

Proof. We start with an informal description of the proof. The heart of the argument is showing that the cuts that participate in a supposedly good ℓ_1 embedding of X cannot be balanced on one hand, and cannot be imbalanced on the other. First notice that the average distance in X is almost double that of the distance between $\mathbf{0}$ and any other point (achieving this in a cube structure without violating the triangle inequality was where the tensor operation came in handy). For a cut metric on the points of X, such a relation occurs only for very imbalanced cuts; hence the representation of balanced cuts in a low distortion embedding cannot be large. On the other hand, comparing the (overall) average distance to the average distance between neighboring points in the cube shows that any good embedding must use cuts with very small edge expansion, and such cuts in the cube must be balanced (the same argument says that one must use the dimension cuts when embedding the hamming cube into ℓ_1 with low distortion). The fact that only *symmetric cuts* participate in the ℓ_1 embedding (or else the distortion becomes infinite due to the tensor operation) enables us to use the stronger isoperimetric inequality which leads to the current lower bound. We now proceed to the proof.

We may view X as a distance function with points in $\mathbf{u}\in Q_n\cup\{\mathbf{0}\}$, and $d(\mathbf{u},\mathbf{v})=\|\tilde{\mathbf{u}}-\tilde{\mathbf{v}}\|^2$. We first notice that X is indeed a metric space, i.e., that triangle inequalities are satisfied: notice that $X\setminus\{\mathbf{0}\}$ is a subset of $\{-1,1\}^{n^2}$. Therefore, the square Euclidean distances is the same (upto a constant) as their ℓ_1 distance. Hence, the only triangle inequality we need to check is $\|\tilde{\mathbf{u}}-\tilde{\mathbf{v}}\|^2\leq\|\tilde{\mathbf{u}}-\mathbf{0}\|^2+\|\tilde{\mathbf{v}}-\mathbf{0}\|^2$, which is implied by the fact that $\tilde{\mathbf{u}}\cdot\tilde{\mathbf{v}}=(\mathbf{u}\cdot\mathbf{v})^2$ is always nonnegative.

For every $\mathbf{u},\mathbf{v}\in Q_n$, we have $d(\mathbf{u},\mathbf{0})=\|\tilde{\mathbf{u}}\|^2=\tilde{\mathbf{u}}\cdot\tilde{\mathbf{u}}=(\mathbf{u}\cdot\mathbf{u})^2=n^2$, and $d(\mathbf{u},\mathbf{v})=\|\tilde{\mathbf{u}}-\tilde{\mathbf{v}}\|^2=\|\tilde{\mathbf{u}}\|^2+\|\tilde{\mathbf{v}}\|^2-2(\tilde{\mathbf{u}}\cdot\tilde{\mathbf{v}})=2n^2-2(\mathbf{u}\cdot\mathbf{v})^2$. In particular, if $\mathbf{uv}\in E$ we have $d(\mathbf{u},\mathbf{v})=2n^2-2(n-2)^2=8(n-1)$. We next notice that

$$\sum_{\mathbf{u},\mathbf{v}\in Q_n} d(\mathbf{u},\mathbf{v})=2^{2n}\times 2n^2-2\sum_{\mathbf{u},\mathbf{v}}(\mathbf{u}\cdot\mathbf{v})^2=2^{2n}\times 2n^2-2\sum_{\mathbf{u},\mathbf{v}}\left(\sum_i \mathbf{u}_i\mathbf{v}_i\right)^2=2^{2n}(2n^2-2n),$$

as $\sum_{\mathbf{u},\mathbf{v}}\mathbf{u}_i\mathbf{v}_i\mathbf{u}_j\mathbf{v}_j$ is 2^{2n} when $i=j$, and 0 otherwise.

Let f be a nonexpanding embedding of X into ℓ_1. Notice that

$$d(\mathbf{u},-\mathbf{u})=2n^2-2(\mathbf{u}\cdot\mathbf{v})^2=0,$$

and so any embedding with finite distortion must satisfy $f(\mathbf{u}) = f(-\mathbf{u})$. Therefore inequality (3) can be used and we get that

$$\frac{\alpha \sum_{\mathbf{uv}\in E} \|f(\tilde{\mathbf{u}}) - f(\tilde{\mathbf{v}})\|_1 + \frac{1}{2}\sum_{\mathbf{u}\in Q_n} \|f(\tilde{\mathbf{u}}) - f(\mathbf{0})\|_1}{\frac{1}{2^n}\sum_{\mathbf{u},\mathbf{v}\in Q_n} \|f(\tilde{\mathbf{u}}) - f(\tilde{\mathbf{v}})\|_1} \geq \frac{4}{7}(4\alpha + 1/2). \tag{16}$$

On the other hand,

$$\frac{\alpha \sum_{\mathbf{uv}\in E} d(\mathbf{u},\mathbf{v}) + \frac{1}{2}\sum_{\mathbf{u}\in Q_n} d(\mathbf{u},\mathbf{0})}{\frac{1}{2^n}\sum_{\mathbf{u},\mathbf{v}\in Q_n} d(\mathbf{u},\mathbf{v})} = \frac{8\alpha(n^2-n)+n^2}{4n^2-4} = \frac{1}{2}(4\alpha+1/2) + o(1). \tag{17}$$

The discrepancy between (16) and (17) shows that for every $\epsilon > 0$ and for sufficiently large n, the required distortion of V into ℓ_1 is at least $8/7 - \epsilon$. ∎

6. Discussion. We have considered the metric characterization of SDP relaxations of Vertex Cover and specifically related the amount of "ℓ_1 information" that is enforced with the resulting integrality gap. We showed that no integrality gap exists in the most powerful extreme, i.e., when ℓ_1 embeddability of the solution is enforced. We further demonstrated that integrality gap is not a continuous function of the possible distortion that is allowed, as it jumps from 1 to $2 - o(1)$ when the allowed distortion changes from 1 to $1 + \delta$.

The natural extensions of these results are as follows: (i) check whether the addition of more k-gonal inequalities (something that can be done efficiently for any finite number of such inequalities) can reduce the integrality gap or prove otherwise. It is interesting to note that related questions are discussed in the context of LP relaxations of Vertex Cover and Max Cut in [3, 11]; (ii) use the nonembeddability construction and technique in section 5 to find negative type metrics that incur more significant distortion when embedded into ℓ_1. After the completion of this work, point (i) above was partially resolved [14], as it was shown that the integrality gap remains $2-o(1)$ even when all k-gonal inequalities with $k = O(\sqrt{\log n/\log\log n})$ are added to the standard SDP. It is also important to understand our results in the context of the Lift and Project system defined by Lovász and Schrijver [24], specifically the one that uses positive semidefinite constraints, called LS_+ (see [2] for relevant discussion). A new result of Georgiou, Magen, Pitassi, and Tourlakis [13] shows that after a superconstant number of rounds of LS_+, the integrality gap is still $2-o(1)$. Such results are related, however incomparable in general, to Theorem 5. For more related discussion we refer the reader to [14].

Last, we suggest looking at connections of ℓ_1-embeddability and integrality gaps for other NP-hard problems. Under certain circumstances, such connections may be used to convert hardness results of combinatorial problems into hardness results of approximating ℓ_1 distortion.

Acknowledgment. Special thanks to George Karakostas for very valuable discussions.

REFERENCES

[1] A. Agarwal, M. Charikar, K. Makarychev, and Y. Makarychev, $O(\sqrt{\log n})$ *approximation algorithms for min UnCut, min* 2*CNF deletion, and directed cut problems*, in STOC '05: Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing, ACM Press, New York, 2005, pp. 573–581.

[2] S. Arora, M. Alekhnovich, and I. Tourlakis, *Towards strong nonapproximability results in the Lovász-Schrijver hierarchy*, in STOC '05: Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing, ACM, New York, 2005.

[3] S. Arora, B. Bollobas, L. Lovász, and I. Tourlakis, *Proving integrality gaps without knowing the linear program*, Theory Comput., 2 (2006), pp. 19–51.

[4] S. Arora, J. Lee, and A. Naor, *Euclidean distortion and the sparsest cut [extended abstract]*, in STOC'05: Proceedings of the 37th Annual ACM Symposium on Theory of Computing, ACM, New York, 2005, pp. 553–562.

[5] S. Arora, S. Rao, and U. Vazirani, *Expander flows, geometric embeddings and graph partitioning*, in Proceedings of the 36th Annual ACM Symposium on Theory of Computing, ACM, New York, 2004, pp. 222–231 (electronic).

[6] M. Charikar, *On semidefinite programming relaxations for graph coloring and vertex cover*, in SODA '02: Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 2002, pp. 616–620.

[7] S. Chawla, A. Gupta, and H. Räcke, *Embeddings of negative-type metrics and an improved approximation to generalized sparsest cut*, in SODA '05: Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Vancouver, BC, Canada, 2005, pp. 102–111.

[8] N. Devanur, S. Khot, R. Saket, and N. Vishnoi, *Integrality gaps for sparsest cut and minimum linear arrangement problems*, in Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing, 2006.

[9] M. Deza and M. Laurent, *Geometry of Cuts and Metrics*, Springer-Verlag, Berlin, 1997.

[10] I. Dinur and S. Safra, *On the hardness of approximating minimum vertex-cover*, Ann. Math., 162 (2005), pp. 439–486.

[11] W. Fernandez de la Vega and C. Kenyon-Mathieu, *Linear programming relaxations of maxcut*, in Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms, 2007.

[12] P. Frankl and V. Rödl, *Forbidden intersections*, Trans. Amer. Math. Soc., 300 (1987), pp. 259–286.

[13] K. Georgiou, A. Magen, T. Pitassi, and I. Tourlakis, *Integrality gaps of* 2 - $o(1)$ *for vertex cover sdps in the Lovész-Schrijver hierarchy*, in FOCS, 2007, pp. 702–712.

[14] K. Georgiou, A. Magen, and I. Tourlakis, *Vertex cover resists SDPs tightened by local hypermetric inequalities*, in Proceedings of the 13th Conference on Integer Programmong and Combinatorial Optimization (IPCO 2008), pp. 140–153.

[15] M. X. Goemans and D. P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. Assoc. Comput. Mach., 42 (1995), pp. 1115–1145.

[16] E. Halperin, *Improved approximation algorithms for the vertex cover problem in graphs and hypergraphs*, SIAM J. Comput., 31 (2002), pp. 1608–1623.

[17] H. Hatami, A. Magen, and E. Markakis, *Integrality gaps of semidefinite programs for vertex cover and relations to* ℓ_1 *embeddability of negative type metrics*, in APPROX-RANDOM, 2007, pp. 164–179.

[18] G. Karakostas, *A better approximation ratio for the vertex cover problem*, in Proceedings of the Thirty-Second International Colloquium on Automata, Languages and Programming, 2005.

[19] S. Khot, *On the power of unique* 2*-prover* 1*-round games*, in Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing, ACM, New York, 2002, pp. 767–775.

[20] S. Khot and O. Regev, *Vertex cover might be hard to approximate to within* $2-\epsilon$, in Proceedings of the 18th IEEE Conference on Computational Complexity, 2003, pp. 379–386.

[21] S. Khot and N. Vishnoi, *The unique games conjecture, integrality gap for cut problems and embeddability of negative type metrics into* ℓ_1, in Proceedings of The 46th Annual Symposium on Foundations of Computer Science, 2005.

[22] J. Kleinberg and M. X. Goemans, *The Lovász theta function and a semidefinite programming relaxation of vertex cover*, SIAM J. Discrete Math., 11 (1998), pp. 196–204.

[23] R. Krauthgamer and Y. Rabani, *Improved lower bounds for embeddings into* l_1, in Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 2006.

[24] L. Lovász and A. Schrijver, *Cones of matrices and set-functions and* 0-1 *optimization*, SIAM J. Optim., 1 (1991), pp. 166–190.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 195–204

DISJOINT COLOR-AVOIDING TRIANGLES*

RAPHAEL YUSTER†

Abstract. A set of pairwise edge-disjoint triangles of an edge-colored K_n is *r-color avoiding* if it does not contain r monochromatic triangles, each having a different color. Let $f_r(n)$ be the maximum integer so that in every edge coloring of K_n with r colors, there is a set of $f_r(n)$ pairwise edge-disjoint triangles that is r-color avoiding. We prove that $0.1177n^2(1-o(1)) < f_2(n) < 0.1424n^2(1+o(1))$. The proof of the lower bound uses probabilistic arguments, fractional relaxation and some packing theorems. We also prove that $f_r(n)/n^2 < \frac{1}{6}(1-0.145^{r-1}) + o(1)$. In particular, for every r, if n is sufficiently large, there are edge colorings of K_n with r colors so that the removal of any $o(n^2)$ members from any Steiner triple system does not turn it r-color avoiding.

Key words. edge coloring, packing, triangles

AMS subject classifications. 05C15, 05C35, 05C70

DOI. 10.1137/060666664

1. Introduction.

All graphs considered here are finite, undirected, and simple. For standard graph-theoretic terminology the reader is referred to [1]. The study of properties of edge colorings of K_n is a central topic of research in Ramsey theory and extremal graph theory. In this paper a coloring always refers to an *edge* coloring.

A subgraph of a colored K_n is *monochromatic* if all of its edges are colored with the same color. A set of pairwise edge-disjoint subgraphs of a colored K_n is *r-color avoiding* if it does not contain r monochromatic elements, each having a different color. For an r-coloring C of K_n, and for an integer $k \geq 3$, let $f_{r,k}(C)$ be the maximum size of a set of pairwise edge-disjoint copies of K_k in K_n that is r-color avoiding. Let $f_{r,k}(n)$ be the minimum possible value of $f_{r,k}(C)$, where C ranges over all r-colorings of K_n. When $k=3$, we denote $f_{r,k}(C) = f_r(C)$ and $f_{r,k}(n) = f_r(n)$. Thus, the value $f_2(n)$ guarantees that in *any* red-blue coloring of K_n we will always have a set of $f_2(n)$ edge-disjoint triangles that either does not contain a blue triangle or else does not contain a red one. The main result of this paper establishes nontrivial lower and upper bounds for $f_2(n)$.

THEOREM 1.1.

$$0.1177 - o(1) < \frac{f_2(n)}{n^2} < \frac{3\sqrt{5}-5}{12} + o(1).$$

Notice that $(3\sqrt{5}-5)/12 < 0.1424$. The term $o(1)$ denotes a quantity that tends to 0 as $n \to \infty$. The constant 0.1177 in the lower bound in Theorem 1.1 may be taken to be $(3\beta^2 - \beta^4)/12$, where $\beta = 0.7648\ldots$ is the smallest root of $x^4 - 3x^3 + 1$. Multiplying the constants by 600, we obtain that, in terms of covering percentages, we can always cover more than 70% of the edges with a set of triangles that is 2-color avoiding, while we cannot, in general, expect to cover more than 86% of the edges with such a set. The main difficulty in the proof of Theorem 1.1 is in the lower bound. Our proof for it requires the use of some probabilistic arguments, some known packing theorems, and the use of fractional relaxation and a connection between it and the

*Received by the editors August 2, 2006; accepted for publication (in revised form) July 28, 2008; published electronically December 17, 2008.
http://www.siam.org/journals/sidma/23-1/66666.html
†Department of Mathematics, University of Haifa, Haifa 31905, Israel (raphy@math.haifa.ac.il).

integral problem. Closing the gap between the upper and lower bounds in Theorem 1.1 is currently beyond our reach.

The upper bound follows from a general construction. Notice that a $(\frac{1}{6}-o(1))n^2$ upper bound for $f_r(n)$ is trivial since every set of pairwise edge-disjoint triangles (we also use the expression *triangle packing*) has at most $n(n-1)/6$ elements. In fact, it is well known that $f_r(n) = (\frac{1}{6}-o(1))n^2$ if r is sufficiently large as a function of n, as Kirkman [9] proved that there are triangle packings with $\frac{n^2}{6}(1-o(1))$ triangles. Our construction, however, shows that no finite number of colors suffices to guarantee an asymptotic optimal r-color avoiding triangle packing for *all* n.

THEOREM 1.2. *For all* $r \geq 2$, $\frac{f_r(n)}{n^2} < \frac{1}{6}(1-\zeta^{r-1}) + o(1)$, *where* $\zeta = \frac{7-3\sqrt{5}}{2} > 0.145$.

We briefly mention three related parameters that have been investigated by several researchers. Erdős et al. [4] considered the function $N(n,k)$, which is the minimum number of pairwise edge-disjoint monochromatic K_k in any 2-coloring of K_n. Erdős conjectured that $N(n,3) = n^2/12+o(n^2)$. This conjecture is still open. A lower bound of slightly more than $n^2/13$ is given in [8]. Similarly, let $N'(n,k)$ be the minimum number of pairwise edge-disjoint monochromatic K_k, all in the same color, in any 2-coloring of K_n. Jacobson (see, e.g., [4]) conjectured that $N'(n,3) = n^2/20 + o(n^2)$ (there is a simple example showing this would be the best possible). Again, the result from [8] immediately implies a lower bound of slightly more than $n^2/26$. For a fixed graph H and a 2-coloring C of K_n, let $f_H(C)$ be the number of edges that do not belong to monochromatic copies of H. Now let $f(n,H) = \max_C f_H(C)$. It is shown in [7] that if H is a complete graph (or, in fact, any edge-color-critical graph) and n is sufficiently large, then $f(n,H)$ equals the Turán number $ex(n,H)$.

The rest of this paper is organized as follows. The proof of the lower bound in Theorem 1.1 is given in section 2. The proof of the general upper bound yielding Theorem 1.2 is given in section 3. Notice that the case $r = 2$ of Theorem 1.2 coincides with the upper bound in Theorem 1.1. In section 4 we give some nontrivial proofs of the exact value of $f_2(n)$ for $n \leq 8$. The final section contains some concluding remarks and open problems.

2. A lower bound for $f_2(n)$. The proof of the lower bound in Theorem 1.1 is obtained by combining two different approaches; one approach (which we call the quadratic approach) is more suitable for colorings where no color is significantly more frequent than the other, and the second approach (the fractional approach) is more suitable when one color is significantly more frequent than the other.

For an integer $k \geq 3$, a Steiner system $S(2,k,n)$ is a set X of n points, and a collection of subsets of X of size k (called blocks), such that any 2 points of X are in exactly one of the blocks. In the case $k = 3$, we have a Steiner triple system, which exists if and only if $n \equiv 1,3 \bmod 6$. The case $k = 4$ is known to exist if and only if $n \equiv 1,4 \bmod 12$; see, e.g., [2].

In the proof of the lower bound for Theorem 1.1 we assume that C is a red-blue coloring of K_n with $\alpha\binom{n}{2}$ blue edges and $(1-\alpha)\binom{n}{2}$ red edges, and $1/2 \leq \alpha \leq 1$. We will also assume that $n \equiv 1 \bmod 12$ as this does not affect the asymptotic results. Each approach will yield a lower bound for $f_2(n)$ in terms of n and α. For each plausible α, one of these lower bounds will be at least as large as the claimed lower bound in Theorem 1.1.

2.1. The quadratic approach. For a red-blue coloring C of K_n, let $t(C)$ be the number of monochromatic triangles. Let $t(n,m)$ be the maximum value of $t(C)$

ranging over colorings with m blue edges. Clearly, $t(n,m) = t(n, \binom{n}{2} - m) = \Theta(n^3)$. Goodman [5] conjectured the value of $t(n,m)$. This conjecture has been proved by Olpp [10], who determined $t(n,m)$, and also determined at least one coloring with m blue edges having $t(n,m)$ monochromatic triangles.

Before we state Olpp's result we need to define two graphs. Let u and v be two integers which satisfy $m = \binom{v}{2} + u$, where $0 \le u \le v-1$. Note that for every $m \ge 0$, v and u are uniquely defined. Let $H_1(n,m)$ be the n-vertex graph which is composed of a clique on v vertices and, if $u > 0$, a unique vertex outside the clique, which is connected to exactly u vertices of that clique. (The remaining vertices, if there are any, are isolated.) Note that $H_1(n,m)$ has exactly m edges. Let $H_2(n,m)$ be the complement of $H_1(n, \binom{n}{2} - m)$. Note that $H_2(n,m)$ has exactly m edges. Olpp has proved the following lemma.

LEMMA 2.1 (see Olpp [10]). *Let C_1 be the coloring of K_n, where the edges colored blue are defined by $H_1(n,m)$. Let C_2 be the coloring of K_n, where the edges colored blue are defined by $H_2(n,m)$. Then $t(n,m) = \max\{t(C_1),\ t(C_2)\}$.*

Note that Lemma 2.1 also supplies a formula for $t(n,m)$ since $t(C_1)$ and $t(C_2)$ can be explicitly computed.

LEMMA 2.2. *If C is a red-blue coloring with $m = \alpha\binom{n}{2}$ blue edges and $\alpha \ge 0.5$, then*

$$f_2(C) \ge \frac{n^2}{12}(1 + 3\alpha(1 - \sqrt{\alpha})) - o(n^2).$$

Proof. Let C_1 and C_2 be the colorings in Lemma 2.1, where $m = \alpha\binom{n}{2}$. By examining the graphs $H_1(n,m)$ and $H_2(n,m)$ it is easy to verify that

$$t(C_1) = \binom{n}{3}(1 - 3\alpha(1 - \sqrt{\alpha})) - o(n^3),$$
$$t(C_2) = \binom{n}{3}(1 - 3(1-\alpha)(1 - \sqrt{1-\alpha})) - o(n^3).$$

Since $\alpha \ge 0.5$, we have $t(C_1) \ge t(C_2)$. Thus, by Lemma 2.1,

$$t(C) \le t(n,m) = \binom{n}{3}(1 - 3\alpha(1 - \sqrt{\alpha})) - o(n^3). \tag{2.1}$$

Fix a Steiner triple system $S(2,3,n)$. A random permutation π of $[n]$ that maps the vertices of K_n to the elements of $S(2,3,n)$ corresponds to a random triangle packing L_π of K_n of order $n(n-1)/6$. Every triangle is equally likely to appear in L_π, each with probability $1/(n-2)$. The expected number of monochromatic triangles in L_π is, therefore, equal to $t(C)/(n-2)$. Fix a π for which L_π contains at most $t(C)/(n-2)$ monochromatic triangles. Thus, there is a packing $M \subset L_\pi$, of size at least $|L_\pi| - t(C)/(2n-4)$ which is 2-color avoiding. By (2.1),

$$\begin{aligned} f_2(C) &\ge \frac{n(n-1)}{6} - \frac{t(C)}{2n-4} \\ &\ge \frac{n(n-1)}{6} - \frac{1}{2n-4}\left(\binom{n}{3}(1 - 3\alpha(1 - \sqrt{\alpha})) - o(n^3)\right) \\ &\ge \frac{n^2}{12}(1 + 3\alpha(1 - \sqrt{\alpha})) - o(n^2). \qquad \square \end{aligned}$$

2.2. The fractional approach. We start with the definition of our fractional relaxation. For a red-blue coloring C of K_n, let $\mathcal{T}_r$ be the set of triangles that contain a red edge and let $\mathcal{T}_b$ be the set of triangles that contain a blue edge. A *fractional* blue-avoiding packing is a function $\nu : \mathcal{T}_r \to [0,1]$ satisfying, for each edge e, $\sum_{e \in T \in \mathcal{T}_r} \nu(T) \leq 1$. Similarly, a fractional red-avoiding packing $\nu : \mathcal{T}_b \to [0,1]$ satisfies, for each edge e, $\sum_{e \in T \in \mathcal{T}_b} \nu(T) \leq 1$. The *value* of ν is $|\nu| = \sum_{T \in \mathcal{T}_c} \nu(T)$, where $c = r$ or $c = b$, depending on whether ν is blue-avoiding or red-avoiding. Let $r^*(C)$ (resp., $b^*(C)$) be the maximum possible value of a fractional blue-avoiding (resp., red-avoiding) packing. Let $f_2^*(C) = \max\{r^*(C),\ b^*(C)\}$. Finally, let $f_2^*(n)$ be the minimum of $f_2^*(C)$ ranging over all red-blue colorings of K_n.

It is easy to see that $f_2^*(n) \geq f_2(n)$, by considering only functions ν that take values 0 and 1. It is also not difficult to construct examples showing strict inequality. For example, we trivially have $f_2(4) = 1$, while $f_2^*(4) = 2$. It is interesting, however, and far from trivial, that the gap between $f_2^*(n)$ and $f_2(n)$ cannot be too large. Haxell and Rödl showed in [6] that the gap between a fractional and an integral triangle packing is $o(n^2)$. This, however, is not sufficient since our graphs are colored. In other words, our packings are *not allowed* to assign positive values to certain triangles. In [11] the author has extended the result from [6] to packings whose elements are taken from any given family of graphs, using a different (probabilistic) approach. In fact, the same proof from [11] also holds for *induced* packings. More formally, let $\mathcal{F}$ be any given family of graphs. An *induced* $\mathcal{F}$-packing of a graph G is a set of induced subgraphs of G, each of them isomorphic to an element of $\mathcal{F}$, and any two of them intersecting in at most one vertex. Let $\nu_{\mathcal{F}}(G)$ be the maximum cardinality of an induced $\mathcal{F}$-packing. Similarly, a *fractional* induced $\mathcal{F}$-packing is a function that assigns weights from $[0,1]$ to the induced subgraphs of G that are isomorphic to elements of $\mathcal{F}$, so that for each pair of vertices x, y, the sum of the weights of the subgraphs containing both x and y is at most one. Let $\nu_{\mathcal{F}}^*(G)$ be the maximum value of a fractional induced $\mathcal{F}$-packing.

THEOREM 2.3 (see Yuster [11], induced version). *Let $\mathcal{F}$ be a family of graphs. If G is a graph with n vertices, then $\nu_{\mathcal{F}}^*(G) - \nu_{\mathcal{F}}(G) = o(n^2)$.*

From Theorem 2.3 it is easy to show that $f_2^*(n)$ and $f_2(n)$ are close.

COROLLARY 2.4. $f_2^*(n) - f_2(n) = o(n^2)$.

Proof. Consider a red-blue coloring C of K_n. Let $r(C)$ be the maximum cardinality of a blue-avoiding triangle packing and let $b(C)$ be the maximum cardinality of a red-avoiding triangle packing. It suffices to show that $r^*(C) - r(C) = o(n^2)$ and that $b^*(C) - b(C) = o(n^2)$. Let G be the n-vertex graph obtained by taking only the edges colored red. Consider the family $\mathcal{F} = \{K_3, K_{1,2}, \overline{K_{1,2}}\}$. Clearly, $r(C) = \nu_{\mathcal{F}}(G)$ and $r^*(C) = \nu_{\mathcal{F}}^*(G)$. The result now follows from Theorem 2.3. Similarly $b^*(C) - b(C) = o(n^2)$ by considering the complement of G. ∎

By Corollary 2.4, in order to prove the lower bound claimed for $f_2(n)$ in Theorem 1.1, it suffices to prove the same lower bound for $f_2^*(n)$.

Let $\mathcal{F}_r$ be the set of nonisomorphic graphs on r vertices. We note that each element of $\mathcal{F}_r$ corresponds to a red-blue coloring of K_r by coloring the edges blue and the nonedges red. It is easy to verify that $\mathcal{F}_4$ consists of 11 graphs, each being one of $\{K_4, K_4^-, Q, C_4, P_4, K_{1,3}\}$ or a complement of one of these (the complement of P_4 is P_4; Q is the graph with four edges that contains a triangle). For a graph H let $b^*(H) = b^*(C)$, where C is the red-blue coloring corresponding to H. It is easy to verify that $b^*(K_4) = 2$, $b^*(K_4^-) = 2$, $b^*(Q) = 2$, $b^*(C_4) = 2$, $b^*(P_4) = 2$, $b^*(K_{1,3}) = 1.5$, $b^*(\overline{K_{1,3}}) = 2$, $b^*(\overline{C_4}) = 2$, $b^*(\overline{Q}) = 1.5$, $b^*(\overline{K_4^-}) = 1$, $b^*(\overline{K_4}) = 0$.

LEMMA 2.5. *If C is a red-blue coloring with $m = \alpha\binom{n}{2}$ blue edges and $\alpha \geq 0.5$,*

then

$$f_2(C) \geq \frac{n^2}{12}(3\alpha - \alpha^2) - o(n^2).$$

Proof. By Corollary 2.4 it suffices to prove the claimed lower bound for $f_2^*(C)$. In fact, we shall prove a stronger statement:

$$b^*(C) \geq \frac{n^2}{12}(3\alpha - \alpha^2) - o(n^2). \tag{2.2}$$

Fix a Steiner system $T = S(2,4,n)$ on the set $X = \{1,\ldots,n\}$. We shall also fix, for each block $B = \{i,j,k,\ell\}$ of T, a matching $M(B) = \{\{i,j\},\{k,\ell\}\}$. Let π be a permutation of $[n]$ selected uniformly at random from S_n. The permutation π defines a decomposition of the edges of K_n into a set L_π of $n(n-1)/12$ pairwise edge-disjoint red-blue colored K_4. Indeed, assume that the set of vertices of K_n is $V = \{v_1,\ldots,v_n\}$ and use π to map the blocks of T to pairwise edge-disjoint red-blue colored K_4. A block $B = \{i,j,k,\ell\}$ is mapped to the element of L_π which is the subgraph induced by $\{\pi(i),\pi(j),\pi(k),\pi(\ell)\}$. As noted earlier, each element of L_π corresponds to an element of $\mathcal{F}_4$. Now let

$$f_\pi = \sum_{H \in L_\pi} b^*(H) \leq b^*(C).$$

We will prove that the expectation of the random variable f_π is at least $n^2(3\alpha - \alpha^2)/12 - o(n^2)$, which implies (2.2).

For $H \in \mathcal{F}_4$, let $t_\pi(H)$ denote the number of elements of L_π corresponding to H. Clearly,

$$\sum_{H \in \mathcal{F}_4} t_\pi(H) = \frac{n(n-1)}{12}.$$

We may therefore rewrite f_π as

$$f_\pi = \sum_{H \in \mathcal{F}_4} t_\pi(H) b^*(H). \tag{2.3}$$

We need to estimate the expectation $E[t_\pi(H)]$ for various H.

Our first observation is that $E[t_\pi(K_4)] \leq \frac{\alpha^2}{12} n(n-1)(1-o(1))$. Indeed, consider a block B of T, and consider its preassigned matching $M(B) = \{\{i,j\},\{k,\ell\}\}$. The probability that $(\pi(i),\pi(j))$ is blue is precisely α. The probability that $(\pi(k),\pi(\ell))$ is blue given that we are *told* that $(\pi(i),\pi(j))$ is blue (and even told its identity) is $\alpha(1-o(1))$. Since there are $n(n-1)/12$ blocks we have that $E[t_\pi(K_4)] \leq \frac{\alpha^2}{12} n(n-1)(1-o(1))$. Similarly, $E[t_\pi(\overline{K_4})] \leq \frac{(1-\alpha)^2}{12} n(n-1)(1-o(1))$. However, we can do much better.

LEMMA 2.6.

$$E\left[t_\pi(K_4) + \frac{2}{3}t_\pi(K_4^-) + \frac{1}{3}t_\pi(Q) + \frac{2}{3}t_\pi(C_4) + \frac{1}{3}t_\pi(P_4) + \frac{1}{3}t_\pi(\overline{C_4})\right]$$
$$= \frac{\alpha^2}{12} n(n-1)(1-o(1)).$$
$$E\left[t_\pi(\overline{K_4}) + \frac{2}{3}t_\pi(\overline{K_4^-}) + \frac{1}{3}t_\pi(\overline{Q}) + \frac{2}{3}t_\pi(\overline{C_4}) + \frac{1}{3}t_\pi(P_4) + \frac{1}{3}t_\pi(C_4)\right]$$
$$= \frac{(1-\alpha)^2}{12} n(n-1)(1-o(1)).$$

Proof. For each element $H \in L_\pi$, let $m(H)$ be the number of blue perfect matchings it contains, and let $g_\pi = \sum_{H \in L_\pi} m(H)$. Clearly,

$$g_\pi = 3t_\pi(K_4) + 2t_\pi(K_4^-) + t_\pi(Q) + 2t_\pi(C_4) + t_\pi(P_4) + t_\pi(\overline{C_4}).$$

Since K_4 has precisely three perfect matchings, the expected number of blocks B for which $M(B)$ is mapped to two blue edges is $\frac{1}{3}E[g_\pi]$. On the other hand, the expected number of blocks B for which $M(B)$ is mapped to two blue edges is $\frac{\alpha^2}{12}n(n-1)(1-o(1))$, as noted in the paragraph preceding the lemma. Thus, the first equality in the statement of the lemma follows. The second equality follows analogously. ∎

To simplify notation, consider the following eleven variables: $x_1 = E[t_\pi(K_4)]$, $x_2 = E[t_\pi(K_4^-)]$, $x_3 = E[t_\pi(Q)]$, $x_4 = E[t_\pi(C_4)]$, $x_5 = E[t_\pi(P_4)]$, $x_6 = E[t_\pi(K_{1,3})]$, $x_7 = E[t_\pi(\overline{K_{1,3}})]$, $x_8 = E[t_\pi(\overline{C_4})]$, $x_9 = E[t_\pi(\overline{Q})]$, $x_{10} = E[t_\pi(\overline{K_4^-})]$, $x_{11} = E[t_\pi(\overline{K_4})]$. With these variables, placing expectations on both sides of (2.3) we obtain

$$E[f_\pi] = 2x_1 + 2x_2 + 2x_3 + 2x_4 + 2x_5 + 1.5x_6 + 2x_7 + 2x_8 + 1.5x_9 + x_{10}.$$

Let $y_i = x_i/n(n-1)$ for $i = 1, \ldots, 11$. Using Lemma 2.6, a lower bound for $E[f_\pi]$ is obtained by solving the following linear program:

$$\begin{aligned}
&\min\ 2y_1 + 2y_2 + 2y_3 + 2y_4 + 2y_5 + 1.5y_6 + 2y_7 + 2y_8 + 1.5y_9 + y_{10}\\
&\text{s.t.}\ \sum_{i=1}^{11} y_i = \frac{1}{12},\\
&\quad y_1 + \frac{2}{3}y_2 + \frac{1}{3}y_3 + \frac{2}{3}y_4 + \frac{1}{3}y_5 + \frac{1}{3}y_8 = \frac{\alpha^2}{12} - o(1),\\
&\quad y_{11} + \frac{2}{3}y_{10} + \frac{1}{3}y_9 + \frac{2}{3}y_8 + \frac{1}{3}y_5 + \frac{1}{3}y_4 = \frac{(1-\alpha)^2}{12} - o(1),\\
&\quad y_i \geq 0 \quad \text{for } i = 1, \ldots, 11.
\end{aligned}$$

In order to derive an optimal solution for this linear program, we exhibit matching solutions both for it and for its dual. The dual program is

$$\begin{aligned}
&\max \frac{1}{12}z_1 + \left(\frac{\alpha^2}{12} - o(1)\right)z_2 + \left(\frac{(1-\alpha)^2}{12} - o(1)\right)z_3\\
&\text{s.t.}\\
&\qquad \begin{pmatrix} z_1 & z_2 & z_3 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \frac{2}{3} & \frac{1}{3} & \frac{2}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{2}{3} & \frac{1}{3} & \frac{2}{3} & 1 \end{pmatrix}\\
&\qquad\qquad \leq \begin{pmatrix} 2 & 2 & 2 & 2 & 2 & 1.5 & 2 & 2 & 1.5 & 1 & 0 \end{pmatrix}.
\end{aligned}$$

(In the argument below and, in fact, throughout Lemma 2.6, we could write all expressions explicitly, instead of writing $o(1)$ terms. However, this would be somewhat cumbersome and, moreover, the reader will be able to check that this is not necessary.) A feasible solution for the dual is $z_1 = 3/2$, $z_2 = 1/2$, and $z_3 = -3/2$ (notice that the constraint set of the dual does not involve $o(1)$ terms). The value this solution attains is $(3\alpha - \alpha^2)/12 - o(1)$. To prove that this is, in fact, an asymptotically optimal solution, we exhibit a feasible solution for the primal problem whose value is also $(3\alpha - \alpha^2)/12 - o(1)$. Indeed, consider the solution $y_1 = \alpha^2/12 - o(1)$, $y_{11} = (1-\alpha)^2/12 - o(1)$, and $y_6 = (\alpha - \alpha^2)/6 + o(1)$ and all the other eight variables are zero, so that all constraints are satisfied. Indeed this solution attains the value $(3\alpha - \alpha^2)/12 - o(1)$, as required. It follows that $E[f_\pi] \geq n^2(3\alpha - \alpha^2)/12 - o(n^2)$, as required. ∎

2.3. Combining the results. Given Lemmas 2.2 and 2.5, we see that if $\alpha \geq 0.5$ is close to 0.5 then the bound in Lemma 2.2 is larger than the bound in Lemma 2.5. On the other hand, when α approaches 1, the bound in Lemma 2.5 approaches the optimal packing of size $n^2/6 - o(n^2)$. By equating $1 + 3\alpha(1 - \sqrt{\alpha})$ with $3\alpha - \alpha^2$ we get that the point of equilibrium is the square of the smallest root of $x^4 - 3x^3 + 1$. If $\beta = 0.7648\ldots$ denotes this root we clearly have

$$f_2(n) \geq \frac{3\beta^2 - \beta^4}{12} n^2 - o(n^2) \ ,$$

proving the lower bound in Theorem 1.1. □

3. An upper bound for $f_r(n)$. We start this section with a construction of a red-blue coloring of K_n that cannot avoid a monochromatic red triangle and a monochromatic blue triangle in any large triangle packing.

Let $0 < \alpha < 1$ be a parameter and let A be a set of αn vertices and B a set of $n(1-\alpha)$ vertices. The vertices of A induce a monochromatic red clique, and all other edges are colored blue. Suppose there is a K_3-packing of size x with no monochromatic red K_3. Then, each element of this packing either contains two edges from the cut (A, B) or has all its three vertices from B. Thus,

$$x < \frac{\alpha(1-\alpha)}{2} n^2 + \frac{(1-\alpha)^2}{6} n^2 + o(n^2). \tag{3.1}$$

Suppose there is a packing of size y with no monochromatic blue K_3. Then we cannot use edges with both endpoints in B at all. Thus,

$$y < \frac{\alpha^2/2 + \alpha(1-\alpha)}{3} n^2 + o(n^2). \tag{3.2}$$

Now, let $z = \max\{x, y\}$. By equating (3.1) and (3.2) we get that for $\alpha = (\sqrt{5} - 1)/2$ we have

$$z < \frac{3\sqrt{5} - 5}{12} n^2 + o(n^2) \approx 0.1424 n^2 (1 + o(1)).$$

In particular, this proves the upper bound in Theorem 1.1.

The construction for $r > 2$ generalizes the construction above. Suppose the set of vertices V of K_n is partitioned into vertex classes $V_1, \ldots, V_r$. The edges with both endpoints in V_i are colored with color i, and an edge between V_i and V_j for $i < j$ is colored with color j. The idea is to choose the sizes of the vertex classes so that a sufficiently large K_3-packing must contain an i-monochromatic K_3 for each color i. Fix $0 < \alpha < 1$, and assume that $|V_i| = \alpha(1-\alpha)^{i-1} n$ for $i = 1, \ldots, r-1$ and $|V_r| = (1-\alpha)^{r-1} n$ (we ignore floors and ceilings as these have no effect on the asymptotic result).

Suppose there is a K_3-packing L_i of size x_i with no i-monochromatic K_3. An upper bound for x_1 is identical to the upper bound for x in (3.1):

$$x_1 < \frac{\alpha(1-\alpha)}{2} n^2 + \frac{(1-\alpha)^2}{6} n^2 + o(n^2). \tag{3.3}$$

For $i = 2, \ldots, r-1$, we notice that no two edges inside V_i appear together in a non–i-monochromatic K_3. Since the third vertex of a non–i-monochromatic K_3 having two vertices in V_i must belong to some V_j with $j > i$, we have

$$x_i < \frac{1}{6} n^2 - \frac{\alpha^2(1-\alpha)^{2i-2}}{6} n^2 + \frac{\alpha(1-\alpha)^{i-1}(1-\alpha)^i}{6} n^2 + o(n^2). \tag{3.4}$$

For $i = r$, L_r cannot cover edges with both endpoints in V_r at all. Thus, similarly to (3.2) we get

$$x_r < \frac{(1-(1-\alpha)^{r-1})^2/2 + (1-\alpha)^{r-1}(1-(1-\alpha)^{r-1})}{3} n^2 + o(n^2). \tag{3.5}$$

Simplifying (3.3), (3.4), and (3.5) we get that

$$\frac{6x_i}{n^2} - o(1) \leq 1 - \alpha(2\alpha - 1)(1-\alpha)^{2i-2}, \quad i = 1, \ldots, r-1, \tag{3.6}$$

$$\frac{6x_r}{n^2} - o(1) \leq 1 - (1-\alpha)^{2r-2}. \tag{3.7}$$

Now, let $z = \max\{x_1, \ldots, x_r\}$, and notice that, in fact, it suffices to consider $z = \max\{x_{r-1}, x_r\}$. By equating the case $i = r-1$ in (3.6) with (3.7) we get that for $\alpha = (\sqrt{5}-1)/2$ we have

$$z < \frac{1 - (\frac{3-\sqrt{5}}{2})^{2r-2}}{6} n^2 + o(n^2).$$

It follows that $\frac{f_r(n)}{n^2} < \frac{1}{6}(1-\zeta^{r-1}) + o(1)$, where $\zeta = \frac{7-3\sqrt{5}}{2}$. This completes the proof of Theorem 1.2. □

4. Determining $f_2(n)$ for small n. Clearly, $f_2(3) = f_2(4) = 1$. For $n = 5$ we notice that there are 15 distinct pairs of edge-disjoint triangles. Each of the 10 triangles appears in three of these pairs. If each pair contains a red triangle and a blue triangle we must have five red triangles and five blue triangles. Suppose, w.l.o.g., that there are at most five red edges. Notice that five edges cannot induce five triangles. Thus, $f_2(5) = 2$.

For $n = 6$, notice that K_6 has 15 distinct perfect matchings. Each perfect matching uniquely defines two sets of four pairwise edge-disjoint triangles (by considering the $K_{2,2,2}$ obtained by deleting the matching). All together, there are 30 distinct triangle packings of size 4. Totally, they contain 120 triangles, but since K_6 has 20 triangles, each triangle appears in precisely six such packings. Suppose each packing has a red and a blue monochromatic triangle. Then there are at least five monochromatic red triangles and at least five monochromatic blue triangles. Assume, w.l.o.g., that there are at most seven red edges. Notice that seven edges cannot induce five triangles. It follows that $f_2(6) = 4$. (K_6 does not have five pairwise edge-disjoint triangles.)

For $n = 7$, we first notice that $f_2(7) \leq 6$ (although K_7 does have a Steiner triple system with seven edge-disjoint triangles). Indeed, take a red K_5 and color the remaining 11 edges blue. In a packing that has no red triangle there are at least two blue edges in each triangle, and hence its size is at most 5. In a packing that has no blue triangle the unique blue edge that is not incident with any red edge does not appear. Hence, the packing contains at most six triangles. In fact, it is easy to verify that this coloring indeed contains six edge-disjoint triangles, none of which is entirely blue. For the other direction, K_7 contains precisely 30 distinct Steiner triple systems. Totally, they contain 210 triangles, but since K_7 has 35 triangles, each triangle appears in precisely six such systems. If each system contains two blue triangles and two red triangles, then there are 10 red triangles and 10 blue triangles. Assume, w.l.o.g., that there are at most 10 red edges. The only way 10 edges can induce 10 triangles is if

they form a K_5, and this is precisely the construction we examined earlier. Thus, $f_2(7) = 6$.

For $n = 8$, notice that K_8 has 105 distinct perfect matchings. Each perfect matching uniquely defines eight sets of eight pairwise edge-disjoint triangles (by considering the $K_{2,2,2,2}$ obtained by deleting the matching). All together, there are 840 distinct triangle packings of size 8. Totally, they contain 6720 triangles, but since K_8 has 56 triangles, each triangle appears in precisely 120 such packings. Suppose each packing has two red and two blue monochromatic triangles. Then there are at least 14 monochromatic red triangles and at least 14 monochromatic blue triangles. Assume, w.l.o.g., that there are at most 14 red edges. If there are 14 red edges, then, by Lemma 2.1, $t(8, 14) = 24$, so we cannot have 28 monochromatic triangles. If there are only 13 red edges or fewer, they cannot induce 14 triangles. It follows that $f_2(8) \geq 7$. To see that $f_2(8) = 7$, consider a red K_5 and color the remaining 18 edges blue. In any packing of eight pairwise edge-disjoint triangles, this coloring has both a red and a blue triangle.

5. Concluding remarks. The most obvious open problem is to determine the true asymptotic behavior of $f_2(n)$. We conjecture that the upper bound construction is the right (asymptotic) answer. Namely, $f_2(n) = \frac{3\sqrt{5}-5}{12}n^2 - o(n^2)$. The fractional approach yielding Lemma 2.5 uses the Steiner system $S(2, 4, n)$. At the price of significantly complicating the proof, we can use a higher order system such as $S(2, k, n)$. (Wilson's theorem guarantees the existence of an $S(2, k, n)$ when n is any sufficiently large integer satisfying $n \equiv 1 \bmod k(k-1)$.) This, however, requires the analysis of all possible colorings of K_k and their expected frequencies, which is already a daunting task for $k = 6$, and which will not lead to a significant improvement in the lower bound.

A *Steiner packing* of K_n is a triangle packing of maximum cardinality. As already mentioned, if $n \equiv 1, 3 \bmod 6$, there is a Steiner triple system, which, by definition, is a Steiner packing that covers every edge and hence consists of $n(n-1)/6$ elements. For other moduli, the cardinality of a Steiner packing is also well known [3]. It is $\lfloor n(n-2)/6 \rfloor$ if n is even and $\lfloor n(n-1)/6 - 1 \rfloor$ if $n \equiv -1 \bmod 6$. Let $g(r)$ be the maximum integer n so that in every r-edge coloring of K_n there is a Steiner packing that is r-color avoiding. The arguments in section 4 show that $g(2) = 6$, since already for $n = 7$ the Steiner packing has seven elements while $f_2(7) = 6$. It would be interesting to determine the behavior of $g(r)$ as a function of r. The proof of Theorem 1.2 shows that $g(r)$ is at most exponential in r (the base being at most roughly 2.7).

Acknowledgment. I thank Eli Berger for useful discussions.

REFERENCES

[1] B. Bollobás, *Modern Graph Theory*, Grad. Texts in Math. 184, Springer-Verlag, New York, 1998.

[2] A.E. Brouwer, *Optimal packing of K_4's into a K_n*, J. Combin. Theory Ser. A, 26 (1979), pp. 278–297.

[3] C.J. Colbourn and J.H. Dinitz, *The CRC Handbook of Combinatorial Designs*, CRC Press, Boca Raton, FL, 1996.

[4] P. Erdős, R.J. Faudree, R.J. Gould, M.S. Jacobson, and J. Lehel, *Edge disjoint monochromatic triangles in 2-colored graphs*, Discrete Math., 231 (2001), pp. 135–141.

[5] A.W. Goodman, *Triangles in a complete chromatic graph*, J. Austral. Math. Soc. Ser. A, 39 (1985), pp. 86–93.

[6] P.E. Haxell and V. Rödl, *Integer and fractional packings in dense graphs*, Combinatorica, 21 (2001), pp. 13–38.
[7] P. Keevash and B. Sudakov, *On the number of edges not covered by monochromatic copies of a fixed graph*, J. Combin. Theory Ser. B, 90 (2004), pp. 41–53.
[8] P. Keevash and B. Sudakov, *Packing triangles in a graph and its complement*, J. Graph Theory, 47 (2004), pp. 203–216.
[9] T.P. Kirkman, *On a problem in combinatorics*, Cambridge Dublin Math. J., 2 (1847), pp. 191–204.
[10] D. Olpp, *A conjecture of Goodman and the multiplicities of graphs*, Australas. J. Combin., 14 (1996), pp. 267–282.
[11] R. Yuster, *Integer and fractional packing of families of graphs*, Random Structures Algorithms, 26 (2005), pp. 110–118.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 205–220

ON $(3,1)^*$-COLORING OF PLANE GRAPHS*

BAOGANG XU†

Abstract. Given positive integers k and d, a graph G is said to be $(k,d)^*$-colorable if the vertices of G can be colored with k colors such that every vertex has at most d neighbors receiving the same color as itself. Let $\mathcal{G}$ be the family of plane graphs with neither adjacent triangles nor cycles of length 5. It is proved in this paper that every graph in $\mathcal{G}$ is $(3,1)^*$-colorable. This result is sharp in the sense that there exist non-$(2,1)^*$-colorable plane graphs with neither triangles nor cycles of length 5. As a corollary, after removing a matching, every graph in $\mathcal{G}$ is 3-colorable. This provides a partial solution to a conjecture of Borodin and Raspaud [*J. Combin. Theory Ser. B*, 93 (2003), pp. 17–27].

Key words. triangle, defective coloring, plane graph

AMS subject classifications. 05C15, 05C78

DOI. 10.1137/06066093X

1. Introduction. In 1976, Steinberg conjectured (see [11, p. 229]) that every plane graph without cycles of length 4 and 5 is 3-colorable. Borodin et al. [4] proved that every plane graph without cycles of length from 4 to 7 is 3-colorable. As a variation of Steinberg's 3-coloring problem, Borodin and Raspaud [5] considered the 3-colorability of plane graphs with neither cycles of length 5 nor triangles of shorter distance, proved that every plane graph with neither cycles of length 5 nor triangles of distance less than four is 3-colorable, and proposed a conjecture claiming that every plane graph with neither adjacent triangles nor cycles of length 5 is 3-colorable, where the distance between triangles is the length of the shortest path between vertices of different triangles, and two triangles are said to be adjacent if they have an edge in common.

The result of [5] was improved independently by Borodin and Glebov [3], and Xu [13]. They showed that every plane graph with neither 5-cycles nor triangles of distance less than three is 3-colorable. Xu [14] improved the result of [4] by showing that every plane graph with neither adjacent triangles nor cycles of length 5 and 7 is 3-colorable. Both Steinberg's conjecture and Borodin and Raspaud's conjecture are still open. It seems that there is no expectation to solve these problems completely in the very near future.

In this paper, instead of studying classical colorings, we consider the *defective coloring* problem on plane graphs. Given positive integers k and d, a k-coloring with deficiency d, simply denoted by a $(k,d)^*$-coloring of G, is a mapping $\phi : V(G) \longmapsto \{1,2,\ldots,k\}$ such that every vertex v has at most d neighbors receiving the same color as v itself. A graph is called $(k,d)^*$-*colorable* if it admits a $(k,d)^*$-coloring. A $(k,0)^*$-coloring is just a classical k-coloring of a graph. So, defective coloring is a natural generalization of the classical colorings.

The concept of defective coloring (also called *improper coloring* in some papers) was simultaneously introduced by Burr and Jacobson (see [1]), Cowen, Cowen, and

*Received by the editors May 25, 2006; accepted for publication (in revised form) August 13, 2008; published electronically December 17, 2008. This research was supported by the NSFC.
http://www.siam.org/journals/sidma/23-1/66093.html
†School of Mathematics and Computer Science, Nanjing Normal University, 122 Ninghai Road, Nanjing, 210097, People's Republic of China (baogxu@njnu.edu.cn).

Woodall [6], and Harary and Jones [9]. Defective colorability of plane graphs has been extensively studied since then (see [8], [10], and [12] for more results and references). All plane graphs are $(4,0)^*$-colorable (by the FCT), and all outerplanar graphs are $(3,0)^*$-colorable. It was proved [6] that all plane graphs are $(3,2)^*$-colorable and there exists one that is not $(3,1)^*$-colorable; all outerplanar graphs are $(2,2)^*$-colorable and there exists one that is not $(2,1)^*$-colorable. In [7], the authors proved that the $(2,k)^*$-coloring, for $k \geq 1$, and the $(3,1)^*$-coloring problems are NP-complete even for plane graphs. There is no good characterization for $(2,1)^*$-colorable outerplanar graphs.

Graphs considered in this paper are all finite and simple. Undefined terms can be found in [2]. We use $G = (V, E, F)$ to denote a plane graph with vertex set V, edge set E, and face set F, and use $b(f)$ and $N(f)$ to denote the boundary of a face f and the set of faces adjacent to f, respectively. Two faces are *adjacent* if they share an edge. The degree of a face f, denoted also by $d(f)$, is the length of the facial walk of f. A *k-vertex* (resp., *k-face*) is a vertex (resp., face) of degree k, a *$\leq k$-vertex* (resp., *$\leq k$-face*) is a vertex (resp., face) of degree at most k, and a *$\geq k$-vertex* (resp., *$\geq k$-face*) is defined similarly. An n-face f is called an $(l_1, l_2, \ldots, l_n)$-face if the vertices on $b(f)$ have degree $l_1, l_2, \ldots, l_n$ sequentially. An *m-cycle* (resp., *m-path*) is a cycle (resp., path) with m edges. As usual, a 3-cycle is called a *triangle*. For a subset $S \subset V(G)$, $G \setminus S$ denotes the subgraph of G induced by $V(G) \setminus S$.

Let C be a cycle of a plane graph G. We use $int(C)$ and $ext(C)$ to denote the sets of vertices located inside and outside C, respectively. C is called a *separating cycle* if $int(C) \neq \emptyset \neq ext(C)$, and is called a *nonseparating cycle* otherwise. A *facial cycle* is a nonseparating cycle that is the boundary of a face. For convenience, we still use C to denote the set of vertices of C.

Let C be a separating cycle of G. Suppose that $G \setminus int(C)$ admits a k-coloring ϕ and $G \setminus ext(C)$ admits a k-coloring ψ. If ϕ and ψ coincide on C, then ϕ together with ψ gives a k-coloring of G. This is not true when we consider defective colorings. Even if $G \setminus int(C)$ admits a $(k,1)^*$-coloring ϕ', $G \setminus ext(C)$ admits a $(k,1)^*$-coloring ψ', and ϕ' and ψ' coincide on C, ϕ' and ψ' may not offer any $(k,1)^*$-coloring of G.

For applying the coloring extension method to defective colorings, we introduce a new notion and call it *superextendability*. Let G be a graph, H be an induced subgraph of G, and ϕ be a $(k,1)^*$-coloring of H for some integer $k \geq 2$. The pair (H, ϕ) is said to be *superextendable* in G if ϕ can be extended to a $(k,1)^*$-coloring ϕ' of G such that for every vertex $v \in V(G) \setminus V(H)$ and every neighbor u of v in H, $\phi'(v) \neq \phi(u)$. If (H, ϕ) is superextendable for every $(k,1)^*$-coloring ϕ of H, H is called a *k-superextendable subgraph*.

We use $\mathcal{G}$ to denote the family of plane graphs with neither adjacent triangles nor 5-cycles. In this paper, we focus on the $(3,1)^*$-colorability of graphs in $\mathcal{G}$. So, 3-superextendability is simply referred to as superextendability. Below is our main result.

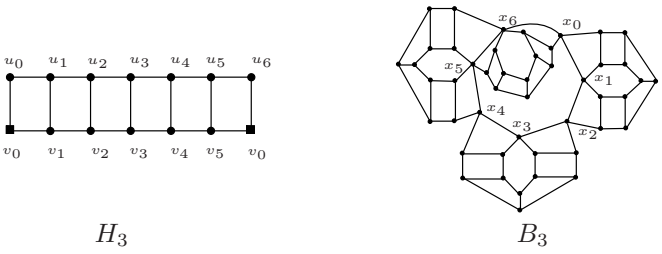
THEOREM 1. *Every triangle or 7-cycle of a graph $G \in \mathcal{G}$ is superextendable.*

Let G be a graph in $\mathcal{G}$. If G has no triangle, then G is 3-colorable by Gröztch's theorem, and is certainly $(3,1)^*$-colorable. If G contains a triangle, then any $(3,1)^*$-coloring of the triangle can be superextended to G. So, as a corollary of Theorem 1, we have the following.

COROLLARY 1. *Every graph in $\mathcal{G}$ is $(3,1)^*$-colorable.*

This result is sharp in the sense that there exist non-$(2,1)^*$-colorable plane graphs with neither triangles nor 5-cycles as evidenced by the following example. Let $l \geq 3$ be an integer, and let H_l be the graph consisting of a cycle $v_0 v_1 \ldots v_{2l-1} v_0$, a path

$u_0u_1\dots u_{2l}$, and edges $\{u_0v_0, u_1v_1, \dots, u_{2l-1}v_{2l-1}, u_{2l}v_0\}$ (see H_3 in Figure 1 as an example, where the two solid squares represent a single vertex v_0). It is obvious that in any $(2,1)^*$-coloring ψ of H_l, $\psi(u_i) = \psi(v_i)$ iff $\psi(u_{i+1}) = \psi(v_{i+1})$ for each $i \in \{0,1,2,\dots,2l-2\}$, and $\psi(u_{2l-1}) = \psi(v_{2l-1})$ iff $\psi(u_{2l}) = \psi(v_0)$. Let $H_l^0, H_l^1, \dots, H_l^l$ be $(l+1)$ copies of H_l, and let B_l be the graph obtained from a cycle $C_{2l+1} = x_0x_1\dots x_{2l}x_0$ and $H_l^0, H_l^1, \dots, H_l^l$ as follows. We first identify v_0u_0 and v_0u_{2l} of H_l^l with $x_{2l}x_{2l-1}$ and $x_{2l}x_0$ of C_{2l+1}, respectively. Then, for $i = 0,1,\dots,l-1$, we identify v_0u_0 and v_0u_{2l} of H_l^i with $x_{2i+1}x_{2i}$ and $x_{2i+1}x_{2i+2}$ of C_{2l+1}, respectively (see B_3 of Figure 1 as an example). Since $l \geq 3$, B_l has neither triangles nor 5-cycles. Assume that B_l admits a $(2,1)^*$-coloring ϕ. We may assume, without loss of generality, that $\phi(x_0) = \phi(x_1)$. From the restriction of ϕ on H_l^0 in B_l, $\phi(x_1) = \phi(x_2)$. Then, $\phi(x_0) = \phi(x_1) = \phi(x_2)$, a contradiction.



H_3 B_3

FIG. 1. *H_3 and B_3.*

Let G be a $(3,1)^*$-colorable graph. Since in any $(3,1)^*$-coloring of G, the edges joining the vertices of the same color form a matching, there exists a matching M in G such that $G - M$ is 3-colorable. As a consequence of Corollary 1, we have the following.

COROLLARY 2. *Every graph $G \in \mathcal{G}$ has matching M such that $G - M$ is 3-colorable.*

This provides a partial solution to a conjecture of Borodin and Raspaud [5].

The rest of the paper is organized as follows. In section 2, we prove several lemmas about a minimum counterexample to Theorem 1. Then we complete the proof of Theorem 1 in section 3 by the discharging method.

2. Structures of the minimum counterexamples. In this section, we always assume that $G \in \mathcal{G}$ is a counterexample to Theorem 1 with minimum $\sigma(G) = |V(G)| + |E(G)|$. Let C be a triangle or a 7-cycle of which a $(3,1)^*$-coloring ϕ cannot be superextended to G, and let $r = |C|$.

If C is a separating cycle, then C is superextendable in both $G \setminus ext(C)$ and $G \setminus int(C)$, and hence is superextendable in G, which contradicts the choice of C. So, we may assume, without loss of generality, that C is the boundary of the outer face f_o of G. Recall that for any cycle C' of G, we still use C' to represent the vertex set of C'.

Let $C' \neq C$ be a triangle or a 7-cycle. If $int(C') \neq \emptyset$, then C is superextendable in $G \setminus int(C')$ and C' is superextendable in $G \setminus ext(C')$, and hence C is superextendable in G. Therefore, we have the following lemma.

LEMMA 1. *G contains neither separating triangles nor separating 7-cycles.*

The next lemma shows that G contains at most one separating 4-cycle, and furthermore, if G contains a separating 4-cycle, then G has some particular structure as depicted below. Let $\mathcal{A} \subseteq \mathcal{G}$ be a subfamily of graphs of which each graph

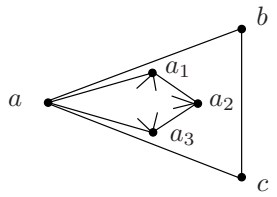


FIG. 2. *The structure of graphs in* $\mathcal{A}$.

has the following properties (see Figure 2): The boundary of its outer face is a triangle, say $abca$, and there exists a particular separating 4-cycle $aa_1a_2a_3a$ such that $ext(aa_1a_2a_3a) = \{b, c\}$.

LEMMA 2. *Suppose that* G *has a separating* 4*-cycle* C_1*. Then,* $G \in \mathcal{A}$*, and* C_1 *is the unique separating* 4*-cycle of* G*.*

Proof. Let $C_1 = v_1v_2v_3v_4v_1$, let $G_1 = G \setminus int(C_1)$, let G_2 be the graph obtained from $G \setminus ext(C_1)$ by substituting $v_1w_1w_2w_3v_2$ for v_1v_2, and let $C_2 = v_1w_1w_2w_3v_2v_3v_4v_1$.

Since $\sigma(G_1) < \sigma(G)$, C is superextendable in G_1 by the minimality of G. Assume that $G \notin \mathcal{A}$. We will show that C_2 is superextendable in G_2. Then C is superextendable in G as C_1 is always superextendable in C_2. This contradicts the choice of C and shows $G \in \mathcal{A}$.

Since $G \in \mathcal{G}$, no edge of C_1 is in any triangles. Therefore, $G_2 \in \mathcal{G}$. To prove that C_2 is superextendable in G_2, we need only check that $\sigma(G_2) < \sigma(G)$. Note that $\sigma(G_2) = \sigma(G \setminus ext(C_1)) + 6$.

If $r = 7$, then $\sigma(C) - \sigma(C \cap C_1) \geq 7$ as $C_1 \neq C$, and thus $\sigma(G_2) = \sigma(G \setminus ext(C_1)) + 6 \leq [\sigma(G) - (\sigma(C) - \sigma(C \cap C_1))] + 6 < \sigma(G)$. So, we assume that $r = 3$.

If $C_1 \cap C = \emptyset$, then $G \setminus int(C_1) - (E(C) \cup E(C_1))$ contains at least one edge (since G is connected), and hence $\sigma(G_2) = \sigma(G \setminus ext(C_1)) + 6 \leq (\sigma(G) - \sigma(C) - 1) + 6 < \sigma(G)$. So, we may further assume that $r = 3$ and $C \cap C_1 \neq \emptyset$.

Since $G \in \mathcal{G}$, $|C \cap C_1| = 1$. If $|ext(C_1)| \geq 3$, then $\sigma(G_2) = \sigma(G \setminus ext(C_1)) + 6 \leq [\sigma(G) - ((\sigma(C) - 1) + 2)] + 6 < \sigma(G)$. Therefore, $|ext(C_1)| = 2$ and $G \in \mathcal{A}$.

If G contains another separating 4-cycle, say C', then C' is a subgraph of $G \setminus ext(C_1)$. By the same arguments as those applied to C_1, one shows that C is superextendable in G. So, C_1 is the unique separating 4-cycle. □

The next lemma shows that $G \setminus C$ contains no 2-vertex, no adjacent 3-vertices, and no 3-face with two 4-vertices and a 3-vertex on its boundary.

LEMMA 3. G *has neither a* ≤ 2*-vertex nor adjacent* 3*-vertices in* $V(G \setminus C)$*, and has no* $(4, 4, 3)$*-face* f *with* $b(f) \cap C = \emptyset$*.*

Proof. If G has a ≤ 2-vertex $v \in V(G \setminus C)$, let $S = \{v\}$. If G has two adjacent 3-vertices u, v in $V(G \setminus C)$, let $S = \{u, v\}$. If G has a $(4, 4, 3)$-face f in $G \setminus C$, let S consist of the three vertices in $b(f)$.

Let $H = G \setminus S$. Then, ϕ has a superextension ϕ_H on H. If (H, ϕ_H) is superextendable in G, then so is (C, ϕ).

If $|S| \leq 2$, then each vertex of S has at most two neighbors in H, and hence ϕ_H is superextendable in G. If $|S| = 3$, let $v_1 \in S$ be a 3-vertex, and let $v_2, v_3 \in S$ be two 4-vertices; then each of v_2 and v_3 has a color not used by ϕ_H on its neighbors in H, and v_1 has two colors not used by ϕ_H on its neighbor in H, and hence ϕ_H is superextendable in G also. Both contradict the choice of C. □

We now show that C is chordless, and for any nonadjacent vertices $x, y \in C$, $N(x) \cap N(y) \subseteq C$.

LEMMA 4. C *is chordless, and for* $x, y \in C$ *with* $xy \notin E(C)$*,* $N(x) \cap N(y) \subseteq C$*.*

Proof. The conclusion is trivial if $r = 3$. So, we suppose that $r = 7$. Let x, y be two vertices on cycle C such that $xy \notin E(C)$. Let P be the shorter path on C joining x and y, and let $l = |E(P)|$. Then, $l = 2$ or 3.

First we assume that $xy \in E(G)$. Since $(C + xy) - E(P)$ contains a $(7 - l + 1)$-cycle, $l = 2$. Assume that $P = xvy$ (see Figure 3(a)). Then, $xvyx$ is a facial cycle (by Lemma 1), and xy is not on any 4-cycle. Let H be the graph obtained from $G \setminus \{v\}$ by inserting a vertex v' into xy (see Figure 3(b), where the broken edges and vertex v are not in H). Then, $H \in \mathcal{G}$, $\sigma(H) = \sigma(G) - 1$, and hence in H, $(C\setminus\{v\})\cup xv'y$ is a superextendable 7-cycle. But this means C itself is superextendable in G, a contradiction to the choice of C. Therefore, $xy \notin E(G)$; i.e., C is chordless.

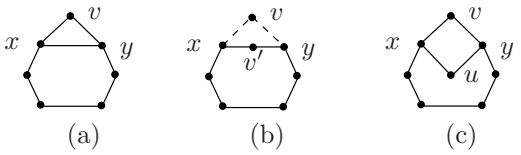


FIG. 3. *Two vertices x and y on C.*

Next, we assume that $(N(x) \cap N(y)) \setminus C$ has a vertex, say u. Since $P \cup xuy$ is an $(l + 2)$-cycle, $l = 2$. Again, let $P = xvy$ (see Figure 3(c)). Since $G \in \mathcal{G}$, $N(u) \cap C = \{x, y\}$. Since $r = 7$, $G \notin \mathcal{A}$, so both $xuyvx$ and $(C \cup xuy) \setminus \{v\}$ are facial cycles (by Lemmas 1 and 2). Thus, $d(u) = 2$ which contradicts Lemma 3. ∎

The next lemma shows that, for an arbitrary 4-face f, $|b(f) \cap C| \leq 2$ and $|b(f) \cap C| = 2$ iff $f \in N(f_o)$ (recall that $b(f)$ and C represent the vertex sets of $b(f)$ and C, respectively).

LEMMA 5. *Suppose that f is a 4-face with $b(f) = v_1v_2v_3v_4v_1$ and $v_1 \in C$. Then, $v_3 \notin C$. Moreover, $|N(v_3) \cap C| = 1$ if $f \in N(f_o)$, and $|N(v_3) \cap C| = 0$ otherwise.*

Proof. Since $G \in \mathcal{G}$, $b(f)$ is chordless. By Lemma 4, $v_3 \in C$ implies that v_2 and v_4 are both in C. So, $v_3 \notin C$.

Suppose first that $f \in N(f_o)$. We suppose, without loss of generality, that $v_2 \in C$. Then, $v_4 \notin C$ (for otherwise, $v_3 \in C$ by Lemma 4). If $N(v_3) \cap C$ has a vertex, say x, other than v_2, then $v_2x \in E(C)$ (by Lemma 4) and hence $v_1v_2xv_3v_4v_1$ is a 5-cycle. Therefore, $N(v_3) \cap C = \{v_2\}$.

Next, we suppose that $f \notin N(f_o)$, and suppose that v_3 has a neighbor, say x, in C. If $r = 3$, let $C = v_1u_1u_2v_1$ and assume that $x = u_2$ (see Figure 4(a)); then $v_1v_2v_3u_2u_1v_1$ is a 5-cycle. So, $r = 7$. Let $C = v_1u_1u_2 \dots u_6v_1$. We may assume that $x \in \{u_4, u_5, u_6\}$ by symmetry. If $x = u_4$, $v_1v_4v_3u_4u_3u_2u_1v_1$ is a separating 7-cycle (see Figure 4(b)). If $x = u_5$, $v_1v_2v_3u_5u_6v_1$ is a 5-cycle (see Figure 4(c)). If $x = v_6$, $v_1v_2v_3u_6v_1$ is a separating 4-cycle that has two common vertices with C (see Figure 4(d)). All are contradictions. Therefore, $|N(v_3) \cap C| = 0$. ∎

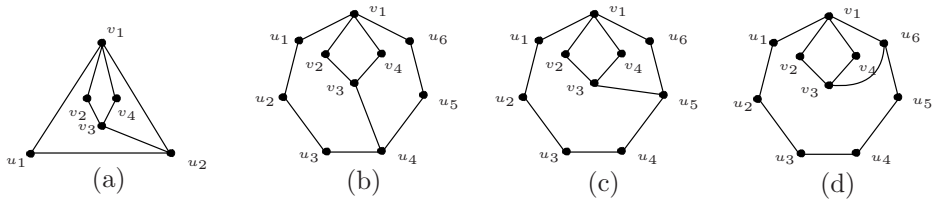


FIG. 4. *$v_1 \in C$ and $N(v_3) \cap C \not\subseteq \{v_2, v_4\}$.*

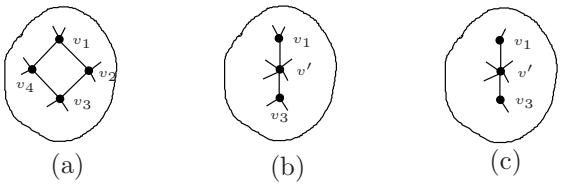


FIG. 5. *Identify a pair of diagonal vertices of a 4-face.*

The following several lemmas are about the distribution of 3-vertices on boundaries of 4-faces in $G \setminus C$. For the convenience of presenting these lemmas, we introduce a new notion. Let $S_1, S_2, \ldots, S_l$ be pairwise disjoint subsets of $V(G)$. We use $G[S_1, S_2, \ldots, S_l]$ to denote the graph obtained from G by identifying all the vertices in S_i to a single vertex for each $i \in \{1, 2, \ldots, l\}$. Lemma 6 below shows that, by identifying a pair of diagonal vertices of a facial 4-cycle, the resulting graph is still in $\mathcal{G}$.

LEMMA 6. *Let f be a 4-face, and let u and v be a pair of diagonal vertices on $b(f)$. Then $G[\{u, v\}] \in \mathcal{G}$.*

Proof. Suppose that $b(f) = uxvyu$. By Lemma 5, we may suppose that $v, y \notin C$.

Since $G \in \mathcal{G}$, G has no 3-path joining u and v, and hence the identification of u and v produces no triangle. If $G[\{u, v\}]$ has a 5-cycle, then G has a 5-path P' joining u and v. If one of x and y is in P', then $b(f) \cup P'$ has a 5-cycle. So, $x, y \notin V(P')$, and hence either $P' \cup uxv$ or $P' \cup uyv$ is a separating 7-cycle; both contradict Lemma 1. Therefore, $G[\{u, v\}] \in \mathcal{G}$. □

This lemma will be used frequently in the proofs of the following four lemmas. The next lemma shows that every 4-face has at most one 3-vertex from $V(G) \setminus C$.

LEMMA 7. *For every 4-face f, $b(f) \setminus C$ contains at most one 3-vertex.*

Proof. Let f be a 4-face with $b(f) = v_1v_2v_3v_4v_1$. Since $G \setminus C$ contains no adjacent 3-vertices by Lemma 3, we may suppose that $f \notin N(f_o)$, and suppose by symmetry that $v_1, v_2, v_3 \notin C$ (see Figure 5(a)). If $d(v_2) = 3$, we are done. So, we suppose that $d(v_2) \geq 4$.

Let $G' = G[\{v_2, v_4\}]$ (see Figure 5(b), where v' is the resulting new vertex), and let $G'' = G' \setminus \{v_1, v_3\}$. By Lemma 6, $G' \in \mathcal{G}$, and so is G''. Since $\sigma(G'') < \sigma(G)$, and C is still a chordless cycle in G'' (by Lemma 5), ϕ has a superextension ψ on G''.

We will show that at least one of v_1 and v_3 is a ≥ 4-vertex. If it is not the case, assume that $d(v_1) = d(v_3) = 3$; then $d_{G'}(v_1) = d_{G'}(v_3) = 2$ (see Figure 5(c)), and so ψ has a superextension ψ' on G' in which $\psi'(v_1) \neq \psi(v') \neq \psi'(v_3)$. Since $d(v_1) = d(v_3) = 3$, $N(v_2) \cap N(v_4) = \{v_1, v_3\}$ (otherwise, assume $u \in (N(v_2) \cap N(v_4)) \setminus \{v_1, v_3\}$ and assume by symmetry that $v_3 \in int(uv_2v_1v_4u)$; then $uv_2v_3v_4u$ is a separating 4-cycle contradicting Lemma 2). By coloring v_2 and v_4 with $\psi(v')$, ψ' yields a $(3, 1)^*$-coloring of G that superextends ϕ and contradicts the choice of C. This contradiction completes the proof of Lemma 7. □

A similar technique will be used in the proofs of the following three lemmas in which we consider the cases where a vertex is incident with more 4-faces.

LEMMA 8. *Let $C_1 = uvv_1u_1u$ and $C_2 = uvv_2u_2u$ be two 4-cycles in $G \setminus C$ (see Figure 6(a)). Then, $d(u_1) = d(u_2) = 3$ only if $d(v) \geq 5$, $d(v_1) = d(v_2) = 3$ only if $d(u) \geq 5$, and for $1 \leq i \neq j \leq 2$, $d(u_i) = d(v_j) = 3$ only if $d(u) \geq 5$ and $d(v) \geq 5$.*

Proof. Since $C_1 \cup C_2 \subseteq V(G \setminus C)$, $u_1v_2, u_2v_1 \notin E(G)$, and C_1 and C_2 are both facial cycles (by Lemma 2). Since $G \in \mathcal{G}$, $\{u_1u_2, u_1v, uv_1, uv_2, u_2v, v_1v_2\} \cap E(G) = \emptyset$.

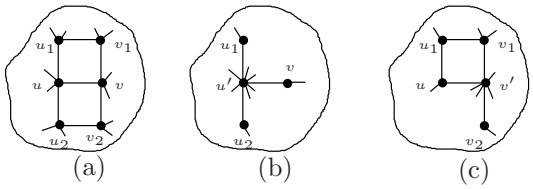


FIG. 6. *Two adjacent* 4*-faces.*

To prove the lemma, we need only show by symmetry that $d(u_1) = d(u_2) = 3$ only if $d(v) \geq 5$, and $d(u_1) = d(v_2) = 3$ only if $d(u) \geq 5$.

First we assume that $d(u_1) = d(u_2) = 3$ and $d(v) \leq 4$. Then, $d(v) = 4$ (by Lemma 7). Let $H_1 = G[\{v_1, u, v_2\}]$, and let u' be the new vertex by identifying u, v_1, and v_2 (see Figure 6(b)).

By Lemma 6, $H_1 \in \mathcal{G}$. Let $H_1' = H_1 \backslash \{u_1, u_2, v\}$. Then, $H_1' \in \mathcal{G}$ and $\sigma(H_1') < \sigma(G)$, and thus ϕ has a superextension $\phi_{H_1'}$ on H_1'. Since $d_{H_1}(u_1) = d_{H_1}(u_2) = d_{H_1}(v) = 2$, $\phi_{H_1'}$ has a superextension ϕ_{H_1} on H_1 in which $\phi_{H_1}(u') \notin \{\phi_{H_1}(u_1), \phi_{H_1}(u_2), \phi_{H_1}(v)\}$. By Lemma 2, no separating 4-cycles may contain vertices in $C_1 \cup C_2$. So, $N(u) \cap N(v_1) = \{u_1, v\}$, $N(u) \cap N(v_2) = \{u_2, v\}$, and $N(v_1) \cap N(v_2) = \{v\}$ (since $d(v) = 4$). By letting $\phi_1(x) = \phi_{H_1}(x)$ for $x \notin \{u, v_1, v_2\}$, and letting $\phi_1(u) = \phi_1(v_1) = \phi_1(v_2) = \phi_{H_1}(u')$, we get a $(3,1)^*$-coloring ϕ_1 of G that superextends ϕ. This contradiction shows that $d(v) \geq 5$ if $d(u_1) = d(u_2) = 3$.

Next, we assume that $d(u_1) = d(v_2) = 3$ and $d(u) \leq 4$. Then, $d(u) = 4$. Let $H_2 = G[\{u_2, v\}]$, let v' be the new vertex in H_2 by identifying u_2 and v (see Figure 6(c)), and let $H_2' = H_2 \setminus \{u_1, u, v_2\}$. Then, $H_2' \in \mathcal{G}$ and $\sigma(H_2') < \sigma(G)$, and hence ϕ has a superextension $\phi_{H_2'}$ on H_2'. Since $d_{H_2}(u_1) = d_{H_2}(u) = 3$ and $d_{H_2}(v_2) = 2$, $\phi_{H_2'}$ has a superextension ϕ_{H_2} such that $\phi_{H_2}(u) \neq \phi_{H_2}(v') \neq \phi_{H_2}(v_2)$. Then, ϕ_{H_2} can be superextended to G since $N(u_2) \cap N(v) = \{u, v_2\}$ (by Lemma 2). This contradiction completes the proof of this lemma. □

Let u be a k-vertex, and let $\{f_0, f_1, \ldots, f_{k-1}\}$ be the set of faces incident with u. We use $H(u)$ to denote the subgraph $\cup_{i=0}^{k-1} b(f_i)$. Figure 7 shows an example of $H(u)$ while $k = 4$.

Our last two lemmas of this section concern the cases where a k-vertex ($k \in \{4, 5\}$) is incident with k 4-faces.

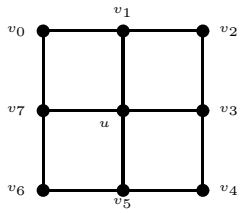


FIG. 7. $H(u)$ *of a* 4*-vertex* u *incident with four* 4*-faces.*

LEMMA 9. *Let* u *be a* 4*-vertex incident with four* 4*-faces and* $V(H(u)) \cap C = \emptyset$. *Suppose that the vertices in* $H(u)$ *are labeled as shown in Figure* 7. *Then,*

(i) $H(u)$ *contains a unique* 3*-vertex if* $\min_{i=0,2,4,6} d(v_i) = 3$*;*

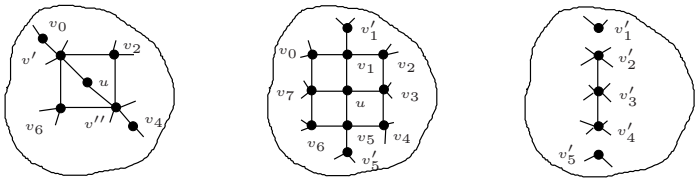
(ii) *for* $j = 1$ *or* 3, $\max\{d(v_j), d(v_{j+4})\} \leq 4$ *only if* $\min\{d(v_j), d(v_{j+4})\} = 3$*; and*

(iii) $N(u)$ *has a* ≥ 5*-vertex, and* $H(u)$ *has at most one* $(4,4,4,3)$*-face.*

Proof. Since $V(H(u)) \cap C = \emptyset$, $H(u)$ contains neither a ≤ 2-vertex nor adjacent 3-vertices (by Lemma 3), and every 4-face in $H(u)$ is incident with at most one 3-vertex (by Lemma 7).

To prove (i), we need only show, by symmetry, that if $d(v_0) = 3$, then all of the other vertices are $\geq$ 4-vertices. Suppose that $d(v_0) = 3$. Then $d(v_1) \geq 4$, $d(v_7) \geq 4$, and $\min\{d(v_2), d(v_3), d(v_5), d(v_6)\} \geq 4$ (by Lemma 8 since $d(u) = 4$). Assume that $d(v_4) = 3$. Let $H_1 = G[\{v_1, v_7\}, \{v_3, v_5\}]$, and let v' and v'' be the new vertices obtained by identifying v_1 with v_7, and v_3 with v_5, respectively (see Figure 8(a)). Then, $d_{H_1}(v_0) = d_{H_1}(u) = d_{H_1}(v_4) = 2$. Let $H_1' = H_1 \setminus \{v_0, u, v_4\}$.

By Lemma 6, $H_1 \in \mathcal{G}$, and so is H_1'. Since $\sigma(H_1') < \sigma(G)$, ϕ has a superextension $\phi_{H_1'}$ on H_1', and $\phi_{H_1'}$ can be superextended to a $(3,1)^*$-coloring ϕ_{H_1} of H_1. By Lemma 2, no separating 4-cycle may contain either $\{v_1, v_7\}$ or $\{v_3, v_5\}$, so $N(v_1) \cap N(v_7) = \{u, v_0\}$ and $N(v_3) \cap N(v_5) = \{u, v_4\}$. By coloring v_1 and v_7 with $\phi_{H_1}(v')$, and coloring v_3 and v_5 with $\phi_{H_1}(v'')$, we get a $(3,1)^*$-coloring of G that superextends ϕ. This contradiction ends the proof of (i).



(a) (b) (c)

FIG. 8. *A 4-vertex u incident with four 4-faces.*

To prove (ii), it suffices to prove that if $d(v_1) \geq 4$ and $d(v_5) \geq 4$, then $d(v_1) \geq 5$ or $d(v_5) \geq 5$. Assume on the contrary that $d(v_1) = d(v_5) = 4$. Let $N(v_1) = \{u, v_0, v_2, v_1'\}$, $N(v_5) = \{u, v_4, v_6, v_5'\}$ (see Figure 8(b)). Let $B = G \setminus \{v_1, u, v_5\}$, and let $B' = B[\{v_0, v_2\}, \{v_3, v_7\}, \{v_4, v_6\}]$. Then, $\sigma(B') < \sigma(G)$. By the similar arguments as used in the proof of Lemma 6, we can show that $B' \in \mathcal{G}$. Then, ϕ has a superextension $\phi_{B'}$ on B'. Let v_2', v_3', and v_4' be the vertices obtained by identifying v_0 with v_2, v_3 with v_7, and v_4 with v_6, respectively (see Figure 8(c)). Since $d(v_1) = d(v_5) = 4$, no separating 4-cycle may contain $\{v_0, v_2\}$, or $\{v_3, v_7\}$, or $\{v_4, v_6\}$ (by Lemma 2), and hence $N_B(v_0) \cap N_B(v_2) = N_B(v_3) \cap N_B(v_7) = N_B(v_4) \cap N_B(v_6) = \emptyset$. By defining $\phi_B(x) = \phi_{B'}(x)$ for $x \in V(B') \setminus \{v_2', v_3', v_4'\}$, and defining $\phi_B(v_0) = \phi_B(v_2) = \phi_{B'}(v_2')$, $\phi_B(v_3) = \phi_B(v_7) = \phi_{B'}(v_3')$, $\phi_B(v_4) = \phi_B(v_6) = \phi_{B'}(v_4')$, ϕ_B is a superextension of ϕ on B. Let $\phi'(x) = \phi_B(x)$ for $x \notin \{u, v_1, v_5\}$, let $\phi'(v_1) \notin \{\phi'(v_0), \phi'(v_1')\}$, $\phi'(v_5) \notin \{\phi'(v_6), \phi'(v_5')\}$, and let $\phi'(u) \notin \{\phi'(v_5), \phi'(v_7)\}$. Then, ϕ' is a superextension of ϕ on G. This contradiction proves (ii).

Now, we proceed to prove (iii). If $d(v_i) = 3$ for some $i \in \{0, 2, 4, 6\}$, then $H(u)$ has a unique 3-vertex (by (i)), and hence has at most one $(4,4,4,3)$-face and has two $\geq$ 5-vertices (by (ii)). So, we suppose that $d(v_i) \geq 4$ for each $i \in \{0, 2, 4, 6\}$. If v_1 is a unique 3-vertex in $H(u)$, then $d(v_3) \geq 5$ or $d(v_7) \geq 5$ (by (ii)), and the conclusion follows immediately. We then suppose, by symmetry, that $d(v_1) = d(v_5) = 3$. Then, by Lemma 8, $d(v_3) \geq 5$ and $d(v_7) \geq 5$, and hence $H(u)$ has no $(4,4,4,3)$-face and has two $\geq$ 5-vertices. □

For convenience, we use V_4 to denote the set of 4-vertices incident with four 4-faces in $G \setminus C$, and use V_4' to denote the subset of vertices of V_4 of which each is incident with a $(4,4,4,3)$-face.

LEMMA 10. *Let u be a 5-vertex incident with five 4-faces such that* $V(H(u)) \cap C = \emptyset$. *Suppose that the vertices in* $H(u)$ *are labeled as shown in Figure 9. Then, for each* $i \in \{1, 3, 5, 7, 9\}$,

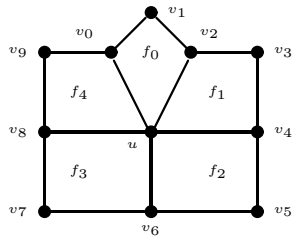


FIG. 9. *$H(u)$ of a 5-vertex u.*

(i) $d(v_i) = d(v_{i+2}) = 3$ *only if* $d(v_j) \geq 4$ *for all* $j \neq i, i+2$. *Moreover,* $d(v_i) = d(v_{i+2}) = 3$ *and* $d(v_{i+3}) = 4$ *only if* $d(v_{i+5}) \notin V_4'$;
(ii) $d(v_i) = d(v_{i+4}) = 3$ *only if* $\max\{d(v_{i+1}), d(v_{i+3})\} \geq 5$; *and*
(iii) $d(v_i) = d(v_{i+5}) = 3$ *only if* $\max\{d(v_{i+1}), d(v_{i+3})\} \geq 5$ *and* $\max\{d(v_{i+7}), d(v_{i+9})\} \geq 5$,

where the summations of subindex are taken modulo 10.

Proof. Since $V(H(u)) \cap C = \emptyset$, $H(u)$ contains neither a ≤ 2-vertex nor adjacent 3-vertices, and every 4-face in $H(u)$ is incident with at most one 3-vertex (by Lemma 7).

If (i) does not hold, we first assume by symmetry that $d(v_1) = d(v_3) = 3$, and $d(v_j) = 3$ for some $j \neq 1, 3$. Then, $j \notin \{0, 2, 4\}$ (by Lemma 7). If $j = 5$, let $H = G[\{v_0, v_2, v_4, v_6\}]$, and let $S = \{u, v_1, v_3, v_5\}$ (note $d_H(w) = 2$ for $w \in S$). If $j = 6$, let $H = G[\{v_0, v_2, v_4\}]$, $S = \{u, v_1, v_3, v_6\}$ (note $d_H(v_1) = d_H(v_3) = 2$ and $d_H(v_6) = d_H(u) = 3$). If $j = 7$, let $H = G[\{v_0, v_2, v_4\}, \{v_6, v_8\}]$, $S = \{u, v_1, v_3, v_7\}$ (then $d_H(w) = 2$ for $w \in S$). If $j = 8$, let $H = G[\{v_0, v_2, v_4\}]$, $S = \{u, v_1, v_3, v_8\}$ (then $d_H(v_1) = d_H(v_3) = 2$ and $d_H(v_8) = d_H(u) = 3$). If $j = 9$, let $H = G[\{v_0, v_2, v_4, v_8\}]$, $S = \{u, v_1, v_3, v_9\}$ (then $d_H(w) = 2$ for $w \in S$).

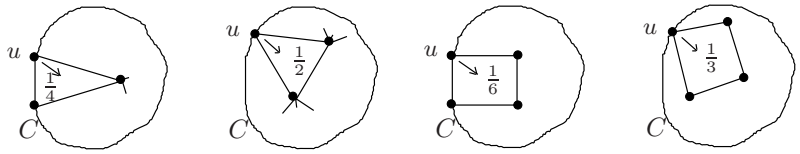
Let $H' = H \setminus S$. By repeatedly applying Lemma 6, we have $H' \in \mathcal{G}$. Since $\sigma(H') < \sigma(G)$, ϕ has a superextension $\phi_{H'}$ on H'. By the degrees of the vertices of S in H, H' is superextendable in H, and any superextension of $\phi_{H'}$ to H yields a superextension of ϕ to G. This contradiction completes the first assertion of (i).

To prove the second assertion of (i), we assume the contrary, by symmetry, that $d(v_1) = d(v_3) = 3$, $d(v_4) = 4$, and $v_6 \in V_4'$. Then, one of x_5, x_6, and x_7 is a 4-vertex adjacent to a 3-vertex $x \in V(G \setminus [C \cup V(H(u))])$. If $x \in N(v_6)$, let $H = G[\{u, v_5\}]$ (then $d_H(v_3) = d_H(v_4) = d_H(v_6) = d_H(x) = 3$), and let $S = \{v_3, v_4, v_6, x\}$. If $d(v_5) = 4$ and $x \in N(v_5)$, let $H = G[\{v_0, v_2, v_4, v_6\}]$ (then $d_H(v_1) = d_H(v_3) = d_H(u) = 2$ and $d_H(v_5) = d_H(x) = 3$), and let $S = \{u, v_1, v_3, v_5, x\}$. If $d(v_7) = 4$ and $x \in N(v_7)$, let $H = G[\{v_0, v_2, v_4\}, \{v_6, v_8\}]$ (then $d_H(v_1) = d_H(v_3) = d_H(u) = 2$ and $d_H(v_7) = d_H(x) = 3$), and let $S = \{u, v_1, v_3, v_7, x\}$. Contradictions can be deduced using the same arguments as above by superextending ϕ to $H \setminus S$, and then to H and G sequentially.

If (ii) does not hold, by symmetry, we may assume that $d(v_1) = d(v_5) = 3$ and $d(v_2) = d(v_4) = 4$. A contradiction can be deduced as previously by letting $H = G[\{v_3, u\}]$ (then $d_H(v_1) = d_H(v_2) = d_H(v_4) = d_H(v_5) = 3$), and letting $S = \{v_1, v_2, v_4, v_5\}$.

If (iii) does not hold, we may assume, by symmetry, that $d(v_1) = d(v_6) = 3$ and $d(v_2) = d(v_4) = 4$. A contradiction can be deduced as previously by letting $H = G[\{v_3, v_5, u\}]$ (then $d_H(v_4) = d_H(v_6) = 2$ and $d_H(v_1) = d_H(v_2) = 3$), and letting $S = \{v_1, v_2, v_4, v_6\}$. □

3. Proof of Theorem 1. We are ready to prove Theorem 1 by the discharging method. Assume that $G \in \mathcal{G}$ is a minimum counterexample to Theorem 1 as described



3-face in $\bar{F}$ 3-face in $\hat{F}$ 4-face in $\bar{F}$ 4-face in $\hat{F}$

FIG. 10. *A vertex u on C and its incident 3- or 4-face.*

in section 2. We define a weight ω on $V(G) \cup F(G)$ by letting $\omega(v) = \frac{d(v)}{3} - 1$ for $v \in V(G)$, and letting $\omega(f) = \frac{d(f)}{6} - 1$ for $f \in F(G)$. Since $\sum_{x \in V(G) \cup F(G)} \omega(x) = \frac{1}{3} \sum_{v \in V(G)} d(v) + \frac{1}{6} \sum_{f \in F(G)} d(f) - |V(G)| - |F(G)|$, by Euler's formula $|V| + |F| - |E| = 2$ of plane graphs, $\sum_{x \in V(G) \cup F(G)} \omega(x) = -2$.

By transferring the weights between elements, we will obtain a new weight ω' on $V(G) \cup F(G)$ such that the total sum of the weights is kept constant while the transferring is in progress. Then the proof will be completed by showing $\sum_{x \in V(G) \cup F(G)} \omega'(x) > -2$. For describing discharging rules, we further introduce some notions. Recall that f_o is the outer face of G.

Let $\bar{F}$ be the set of 3- or 4-faces in $N(f_o)$, and let $\hat{F}$ be the set of 3- or 4-faces f such that $f \notin N(f_o) \cup \{f_o\}$ and $b(f) \cap C \neq \emptyset$. By Lemma 5, $|b(f) \cap C| = 2$ for each face $f \in \bar{F}$, and $|b(f) \cap C| = 1$ for each face $f \in \hat{F}$.

For $i = 3$ or 4, let F_i be the set of i-faces in $G \setminus C$, $F_i' \subset F_i$ be the set of i-faces incident with some 3-vertex, and let $F_{4,j}' \subset F_4'$ be the set of 4-faces incident with $j \geq 5$-vertices for $0 \leq j \leq 3$. Note that $F_4' = \cup_{j=0}^{3} F_{4,j}'$, and $F_{4,0}'$ is just the set of $(4,4,4,3)$-faces in F_4.

It is clear that $\bar{F}, \hat{F}, F_3$, and F_4 are pairwise disjoint, and $\bar{F} \cup \hat{F} \cup F_3 \cup F_4$ contains all 3- and 4-faces but f_o. Recall that $V_4 \subset V(G) \setminus C$ is the set of 4-vertices incident with four 4-faces in F_4, and $V_4' \subset V_4$ is the set of 4-vertices incident with a $(4,4,4,3)$-face.

For $x, y \in V(G) \cup F(G)$, we use $W(x \to y)$ to denote the weight transferred from x to y. Let u be a k-vertex, $v \in N(u)$, and $f \neq f_o$ be an l-face ($3 \leq l \leq 4$) incident with u. The discharging rules are as follows.

(R_1) For $u \in C$, $W(u \to f) = \frac{1}{2l-2}$ if $f \in \bar{F}$, and $W(u \to f) = \frac{1}{l-1}$ if $f \in \hat{F}$ (see Figure 10).

(R_2) For $u \notin C$ and $k \geq 5$,

($R_{2.1}$) $W(u \to f) = \frac{1}{3}$ if $f \in F_3'$, $W(u \to f) = \frac{1}{6}$ if $f \in F_3 \setminus F_3'$ (see Figure 11(a), (b));

($R_{2.2}$) $W(u \to f) = \frac{i+1}{12i}$ if $f \in F_{4,i}'$, $i = 1, 2, 3$ (see Figure 11(c), (d), (e));

($R_{2.3}$) $W(u \to f) = \frac{1}{12}$ if $f \in F_4 \setminus F_4'$ (see Figure 11(f));

($R_{2.4}$) $W(u \to v) = \frac{1}{36}$ if $v \in V_4'$ (see Figure 11(g)).

(R_3) For $u \notin C$ and $k = 4$,

($R_{3.1}$) $W(u \to f) = \frac{1}{6}$ if $f \in F_3$ (see Figure 12(a));

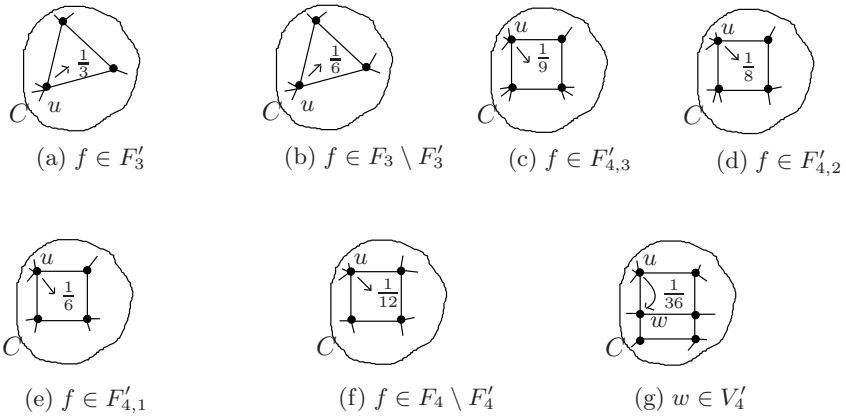
($R_{3.2}$) $W(u \to f) = \frac{1}{9}$ if $f \in F_{4,0}'$ (see Figure 12(b));

($R_{3.3}$) $W(u \to f) = \frac{1}{12}$ if $f \in F_4 \setminus F_{4,0}'$ (see Figure 12(c)).

From the discharging rules, if $u \notin C$ has a neighbor, say v, on C, then u transfers nothing to the two faces incident with uv. Therefore, for each $u \notin C$ with $N(u) \cap C \neq \emptyset$, at least two faces incident with u receive nothing from u.

We turn to calculate ω'.

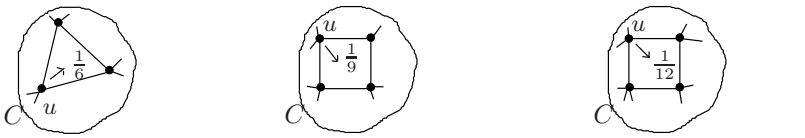
CLAIM 1. *Let $f \neq f_o$ be an l-face. Then, $\omega'(f) \geq 0$, and $\omega'(f) = \omega(f) = \frac{l-6}{6}$ if $l \geq 6$.*



(a) $f \in F_3'$ (b) $f \in F_3 \setminus F_3'$ (c) $f \in F_{4,3}'$ (d) $f \in F_{4,2}'$

(e) $f \in F_{4,1}'$ (f) $f \in F_4 \setminus F_4'$ (g) $w \in V_4'$

FIG. 11. *$A \geq 5$-vertex u and its incident 3- or 4-faces, or its adjacent 4-vertices in V_4'.*



(a) $f \in F_3$ (b) $f \in F_{4,0}'$ (c) $f \in F_4 \setminus F_{4,0}'$

FIG. 12. *A 4-vertex u and its incident 3- or 4-faces.*

Proof. Since G has no 5-cycles, $l \neq 5$. If $l \geq 6$, $\omega'(f) = \omega(f) = \frac{l-6}{6} \geq 0$. Suppose $3 \leq l \leq 4$.

If $f \in \bar{F} \cup \hat{F}$, it totally receives $\frac{1}{l-1}$ from the vertices in $b(f) \cap C$ (by R_1), and then $\omega'(f) = \frac{l-6}{6} + \frac{1}{l-1} = \frac{(l-3)(l-4)}{6(l-1)} = 0$.

If $f \in F_3 \setminus F_3'$, then $b(f)$ has three ≥ 4-vertices of which each transfers $\frac{1}{6}$ to f (by $R_{2.1}$ and $R_{3.1}$). If $f \in F_3'$, then f is incident with a unique 3-vertex and is not a $(4,4,3)$-face (by Lemma 3), so $b(f)$ has a ≥ 5-vertex from which it receives $\frac{1}{3}$ (by $R_{2.1}$), and has a ≥ 4-vertex from which it receives at least $\frac{1}{6}$ (by $R_{2.1}$ and $R_{3.1}$). In either case, $\omega'(f) \geq -\frac{1}{2} + \frac{1}{2} = 0$.

If $f \in F_4 \setminus F_4'$, every vertex on $b(f)$ transfers $\frac{1}{12}$ to f (by $R_{2.3}$ and $R_{3.3}$), so $\omega'(f) = -\frac{1}{3} + 4 \cdot \frac{1}{12} = 0$. Note that $F_4' = \cup_{i=0}^{3} F_{4,i}'$, and each face in F_4' is incident with a unique 3-vertex (by Lemma 7). If $f \in F_{4,i}'$, $1 \leq i \leq 3$, $b(f)$ has $i \geq 5$-vertices and each transfers $\frac{i+1}{12i}$ to f (by $R_{2.2}$) and has $(3-i)$ 4-vertices and each transfers $\frac{1}{12}$ to f (by $R_{3.3}$), so $\omega'(f) = -\frac{1}{3} + i \cdot \frac{i+1}{12i} + (3-i) \cdot \frac{1}{12} = 0$. If $f \in F_{4,0}'$, $b(f)$ has three 4-vertices and each transfers $\frac{1}{9}$ to f (by $R_{3.2}$), then $\omega'(f) = -\frac{1}{3} + 3 \cdot \frac{1}{9} = 0$. ∎

Let u be a k-vertex, and let p_3 and p_4, respectively, be the number of 3-faces and 4-faces in $F_3 \cup F_4$ that are incident with u. Since $G \in \mathcal{G}$, u is incident with at most $\lfloor \frac{k}{2} \rfloor$ 3-faces, and is incident with at least $(p_3 + 1) \geq 6$-faces if $0 < p_3 < \frac{k}{2}$. Therefore,

$$p_3 \leq \left\lfloor \frac{k}{2} \right\rfloor, \text{ and } p_4 \leq \max\{0, k - 2p_3 - 1\} \text{ if } 0 < p_3 < \frac{k}{2}. \tag{1}$$

CLAIM 2. *Let $u \notin C$ be a k-vertex. Then, $\omega'(u) \geq 0$. Moreover, if $d(u) \neq 3$ and $N(u) \cap C \neq \emptyset$, then $\omega'(u) \geq \frac{1}{9}$ unless u is a 5-vertex incident with two 3-faces in F_3'.*

Proof. Since $u \notin C$, $k \geq 3$ by Lemma 3. If $k = 3$, then $\omega'(u) = \omega(u) = 0$.

Suppose that $k = 4$, i.e., $\omega(u) = \frac{1}{3}$. If $N(u) \cap C \neq \emptyset$, then $p_3 \leq 1$ and $p_4 \leq 2 - 2p_3$, and so $\omega'(u) \geq \frac{1}{3} - p_3 \cdot \frac{1}{6} - p_4 \cdot \frac{1}{9} \geq \frac{2+p_3}{18} \geq \frac{1}{9}$ (by R_3). If $N(u) \cap C = \emptyset$ and $p_3 \neq 0$, then $p_4 \leq 2 - p_3$, and so $\omega'(u) \geq \frac{1}{3} - p_3 \cdot \frac{1}{6} - p_4 \cdot \frac{1}{9} \geq 0$ (by R_3). So, we suppose that $N(u) \cap C = \emptyset$ and $p_3 = 0$. If $u \in V_4'$, Lemma 9(iii) ensures that u is incident with a unique $(4, 4, 4, 3)$-face that receives $\frac{1}{9}$ from u (by $R_{3.2}$), and $N(u)$ has a ≥ 5-vertex that transfers $\frac{1}{36}$ to u (by $R_{2.4}$), and then $\omega'(u) \geq \frac{1}{3} - \frac{1}{9} - 3 \cdot \frac{1}{12} + \frac{1}{36} = 0$ as each of the other three 4-faces receives $\frac{1}{12}$ from u (by $R_{3.3}$). If $u \in V_4 \setminus V_4'$, $\omega'(u) = \frac{1}{3} - 4 \cdot \frac{1}{12} = 0$ as u transfers $\frac{1}{12}$ to each face around it (by $R_{3.3}$). If $u \notin V_4$, u is incident with at most three 4-faces and $\omega'(u) \geq \frac{1}{3} - 3 \cdot \frac{1}{9} = 0$ (by $R_{3.2}$).

Next we suppose that $k \geq 5$. Let $q = |N(u) \cap V_4'|$. Then, u transfers totally $\frac{q}{36}$ to its neighbors (by $R_{2.4}$).

We claim that at least $\lceil \frac{q}{2} \rceil$ of the p_4 4-faces incident with u are in $F_4 \setminus F_4'$, of which each receives at most $\frac{1}{12}$ from u (by $R_{2.3}$). Choose an arbitrary $v \in N(u) \cap V_4'$. Then, $V(H(v)) \subseteq G \setminus C$, and v is incident with four 4-faces in F_4 of which one is a $(4, 4, 4, 3)$-face. Suppose $N(v) = \{u, u_1, u_2, u_3\}$ (see Figure 13). If $d(u_6) = 3$, then $v \notin V_4'$ since all vertices in $V(H(u)) \setminus \{u_6\}$ are ≥ 4-vertices (by Lemma 9(i)) and $d(u) \geq 5$. So, $d(u_6) \geq 4$. The same argument shows that $d(u_7) \geq 4$. If $d(u_1) = d(u_3) = 3$, then $d(u_2) \geq 5$ (by Lemma 8) and hence $v \notin V_4'$ too. Therefore, $d(u_1) \geq 4$ or $d(u_3) \geq 4$; at most one of the 4-faces incident with uv is in F_4'. Our claim follows immediately.

By R_2, u transfers at most $\frac{p_3}{3}$ to its incident 3-faces, at most $\frac{1}{6}$ to each of its incident 4-faces in F_4', $\frac{1}{12}$ to each of its incident 4-faces in $F_4 \setminus F_4'$, and $\frac{q}{36}$ to its adjacent 4-vertices in $N(u) \cap V_4'$, $\omega'(u) \geq \frac{k-3}{3} - p_3 \cdot \frac{1}{3} - (p_4 - \lceil \frac{q}{2} \rceil) \cdot \frac{1}{6} - \lceil \frac{q}{2} \rceil \cdot \frac{1}{12} - q \cdot \frac{1}{36} \geq \frac{(k-6)+(k-2p_3-p_4)}{6}$.

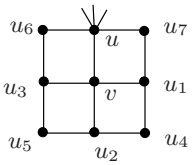


FIG. 13. *A vertex $v \in V_4'$ adjacent to a ≥ 5-vertex u.*

By the second inequality of (1), $k - 2p_3 - p_4 \geq 0$. Furthermore, if $N(u) \cap C \neq \emptyset$, then $p_3 \leq \lceil \frac{k-2}{2} \rceil$ and $p_4 \leq k - 1 - 2p_3$, and hence $k - 2p_3 - p_4 \geq 1$. If $k \geq 6$, then $\omega'(u) \geq \frac{(k-6)+(k-2p_3-p_4)}{6} \geq 0$, and $\omega'(u) \geq \frac{(k-6)+(k-2p_3-p_4)}{6} \geq \frac{1}{6}$ whenever $N(u) \cap C \neq \emptyset$.

If $k = 5$ and $N(u) \cap C \neq \emptyset$, at least two faces incident with u receive nothing from u, $q \leq 2$, and $q = 0$ when u is incident with a 3-face in F_3. Note that u transfers $\frac{1}{3}$ to each of its incident 3-faces in F_3', and transfers at most $\frac{1}{6}$ to each of its incident 4-faces or 3-faces in $F_3 \setminus F_3'$ (by R_2). If u is incident with two 3-faces in F_3', $\omega'(u) \geq \frac{2}{3} - 2 \cdot \frac{1}{3} = 0$ (by $R_{2.1}$). If u is incident with a unique 3-face in F_3', $\omega'(u) \geq \frac{2}{3} - \frac{1}{3} - \frac{1}{6} > \frac{1}{9}$. Otherwise, $\omega'(u) \geq \frac{2}{3} - 3 \cdot \frac{1}{6} - 2 \cdot \frac{1}{36} = \frac{1}{9}$.

Now, the only remaining case is $k = 5$ and $N(u) \cap C = \emptyset$. Since $\omega'(u) \geq \frac{(k-6)+(k-2p_3-p_4)}{6}$, we may further assume that $p_3 = 0$ and $p_4 = 5$ (by (1)). Then, $V(H(u)) \cap C = \emptyset$, and $H(u)$ is as shown in Figure 9. By Lemma 7, $N(u)$ has at most two 3-vertices. We distinguish three cases.

Case 1. $N(u)$ has two 3-vertices. Since every 4-face of $H(u)$ is incident with at most one 3-vertex (by Lemma 7), we may suppose, by symmetry, that $d(v_0) = d(v_6) = 3$ (see Figure 9). Then, $d(v_i) \geq 4$ for $i \in \{1, 2, 4, 5, 7, 8, 9\}$, $d(v_8) \geq 5$ (by applying Lemma 8

to f_3 and f_4), $d(v_2)+d(v_3)\geq 8$, and $d(v_3)+d(v_4)\geq 8$ (by applying Lemma 8 to f_0 and f_1, and to f_1 and f_2, respectively). So,

$$\max\{W(u\to f_3), W(u\to f_4)\}\leq \frac{1}{8} \text{ (by } R_{2.2}).$$

By applying Lemma 9(i) to $H(v_2)$, if $v_2\in V_4$, then v_0 is the unique 3-vertex in $H(v_2)$, so $v_2\notin V_4'$ (since $d(u)=5$). The similar argument shows that $v_4\notin V_4'$ (since $d(u)=5$ and $d(v_6)=3$). So,

$$q=0.$$

If $d(v_3)=3$, then $d(v_2)\geq 5$ (since $d(v_2)+d(v_3)\geq 8$) and $d(v_4)\geq 5$ (since $d(v_2)+d(v_3)\geq 8$), so $W(u\to f_0)\leq \frac{1}{8}$, $W(u\to f_1)\leq \frac{1}{9}$, $W(u\to f_2)\leq \frac{1}{8}$ (by $R_{2.2}$), and so $\omega'(u)\geq \frac{2}{3}-4\cdot\frac{1}{8}-\frac{1}{9}>0$.

If $d(v_3)\geq 4$, then $W(u\to f_1)=\frac{1}{12}$ (by $R_{2.3}$), $W(u\to f_0)\leq \frac{1}{6}$, and $W(u\to f_2)\leq \frac{1}{6}$ (by $R_{2.2}$), and so $\omega'(u)\geq \frac{2}{3}-2\cdot\frac{1}{8}-2\cdot\frac{1}{6}-\frac{1}{12}=0$.

Case 2. $N(u)$ has a unique 3-vertex. By symmetry, we suppose that $d(v_0)=3$ (see Figure 9 also). Then, $d(v_i)\geq 4$ for $i\in\{1,2,4,6,8,9\}$. By Lemma 8, $d(v_2)+d(v_3)\geq 8$ and $d(v_7)+d(v_8)\geq 8$. Since $d(u)=5$ and $d(v_0)=3$, $v_2,v_8\notin V_4'$ (the same arguments as above by Lemma 9(i)), and so $q\leq 2$.

If $d(v_5)=3$, then $d(v_3)\geq 4$ and $d(v_7)\geq 4$ (otherwise, either $d(v_3)=d(v_5)=3$ or $d(v_5)=d(v_7)=3$ implies $d(v_0)\geq 4$ by Lemma 10(i)), and so, $W(u\to f_1)=W(u\to f_3)=\frac{1}{12}$ (by $R_{2.3}$). By applying Lemma 10(iii)) with $i=5$, $d(v_5)=d(v_0)=3$ ensures that $\max\{d(v_2),d(v_4)\}\geq 5$ and $\max\{d(v_6),d(v_8)\}\geq 5$. If $d(v_4)\geq 5$ and $d(v_6)\geq 5$, then $q=0$, and $W(u\to f_2)=\frac{1}{9}$ (by $R_{2.2}$ since $f_2\in F_{4,3}'$), and thus $\omega'(u)\geq \frac{2}{3}-2\cdot\frac{1}{6}-2\cdot\frac{1}{12}-\frac{1}{9}>0$. In each of the remaining three cases ($d(v_2)\geq 5\leq d(v_6)$, or $d(v_2)\geq 5\leq d(v_8)$, or $d(v_4)\geq 5\leq d(v_8)$), there exist two faces among f_0,f_2, and f_4 of which each is incident with two ≥ 5-vertices and receives at most $\frac{1}{8}$ from u (by $R_{2.2}$), and thus $\omega'(u)\geq \frac{2}{3}-\frac{1}{6}-2\cdot\frac{1}{8}-2\cdot\frac{1}{12}-2\cdot\frac{1}{36}>0$.

So, we suppose that $d(v_5)\geq 4$. Then, $W(u\to f_2)=\frac{1}{12}$ (by $R_{2.3}$). If $d(v_3)\geq 4$ and $d(v_7)\geq 4$, then $W(u\to f_1)=W(u\to f_3)=\frac{1}{12}$ (by $R_{2.3}$), and so $\omega'(u)\geq \frac{2}{3}-2\cdot\frac{1}{6}-3\cdot\frac{1}{12}-2\cdot\frac{1}{36}>0$. If $d(v_3)=d(v_7)=3$, then $d(v_2)\geq 5$ and $d(v_8)\geq 5$ (by Lemma 8), each of f_0,f_1,f_3, and f_4 is incident with two ≥ 5-vertices and receives at most $\frac{1}{8}$ from u (by $R_{2.2}$), and thus $\omega'(u)\geq \frac{2}{3}-4\cdot\frac{1}{8}-\frac{1}{12}-2\cdot\frac{1}{36}>0$. If $d(v_3)=3$ and $d(v_7)\geq 4$, then $W(u\to f_3)=\frac{1}{12}$ (by $R_{2.3}$), and $d(v_2)\geq 5$ (by Lemma 8), implying $W(u\to f_0)=W(u\to f_1)=\frac{1}{8}$ (by $R_{2.2}$), and so $\omega'(u)\geq \frac{2}{3}-\frac{1}{6}-2\cdot\frac{1}{8}-2\cdot\frac{1}{12}-2\cdot\frac{1}{36}>0$. The similar argument shows $\omega'(u)>0$ whenever $d(v_3)\geq 4$ and $d(v_7)=3$.

Case 3. $N(u)$ has no 3-vertex. By Lemma 10(i), $H(u)$ has at most two 3-vertices. If $H(u)$ has at most one 3-vertex, then it has four faces, each of which receives $\frac{1}{12}$ from u (by $R_{2.3}$), and so $\omega'(u)\geq \frac{2}{3}-\frac{1}{6}-4\cdot\frac{1}{12}-5\cdot\frac{1}{36}>0$. Therefore, we suppose that $H(u)$ has two 3-vertices, and suppose by symmetry that either $d(v_1)=d(v_3)=3$ or $d(v_1)=d(v_5)=3$ (see Figure 9).

If $d(v_1)=d(v_5)=3$, then $W(u\to f_1)=W(u\to f_3)=W(u\to f_4)=\frac{1}{12}$ (by $R_{2.3}$), and $d(v_2)\geq 5$ or $d(v_4)\geq 5$ (by Lemma 10(ii)) which implies $W(u\to f_0)\leq \frac{1}{8}$ or $W(u\to f_2)\leq \frac{1}{8}$ (by $R_{2.2}$), and so $\omega'(u)\geq \frac{2}{3}-\frac{1}{6}-\frac{1}{8}-3\cdot\frac{1}{12}-4\cdot\frac{1}{36}>0$.

Suppose that $d(v_1)=d(v_3)=3$. Then, $W(u\to f_2)=W(u\to f_3)=W(u\to f_4)=\frac{1}{12}$ (by $R_{2.2}$). By Lemma 8, $v_2\in V_4$ and $d(v_1)=d(v_3)=3$ ensure that v_2 is adjacent to two ≥ 5-vertices, so $v_2\notin V_4'$. Since $d(v_4)=4$ implies $v_6\notin V_4'$ (by taking $i=1$ in Lemma 10(i)), either v_4 or v_6 is not V_4'. So, $q\leq 3$, and $\omega'(u)\geq \frac{2}{3}-2\cdot\frac{1}{6}-3\cdot\frac{1}{12}-3\cdot\frac{1}{36}=0$. This ends the proof of Claim 2. □

Recall that f_o is the outer face of G. All the other faces are called *inner faces*. For convenience, we use $F(u)$ to denote the set of inner faces incident with a vertex u.

CLAIM 3. *Let $u \in C$ be a k-vertex. Then, $\omega'(u) \geq -\frac{1}{3}$. If $F(u)$ has a ≥ 6-face, then $\omega'(u) \geq -\frac{1}{4}$ whenever $k = 3$, and $\omega'(u) \geq -\frac{1}{6}$ whenever $k \geq 4$. Moreover, $k \geq 4$ and $\omega'(u) = -\frac{1}{6}$ only if one face in $N(f_o) \cap F(u)$ is not a 4-face.*

Proof. Only R_1 should be applied in this proof. If $k = 2$, the only face in $F(u)$ is a ≥ 6-face (by Lemma 5), and $\omega'(u) = \omega(u) = -\frac{1}{3}$. If $k = 3$, then $\omega'(u) \geq -2 \cdot \frac{1}{6} = -\frac{1}{3}$ whenever $F(u)$ has two 4-faces, and $\omega'(u) \geq -\frac{1}{4}$ whenever $F(u)$ has a ≥ 6-face. We suppose that $k \geq 4$.

Let t and t' be the number of 3-faces and 4-faces in $F(u)$, respectively. Then, $t' \leq k-1-2t$ if $N(f_o) \cap F(u)$ has a 3-face, and $t' \leq k-1-(2t+1) = k-2-2t$ otherwise. Suppose that the two faces in $N(f_o) \cap F(u)$ have degree k_1 and k_2, respectively.

If $F(u)$ has no ≥ 6-face, then $t = 0$ and $t' \leq k-1$, and so $\omega'(u) = \frac{k-3}{3} - 2 \cdot \frac{1}{6} - (k-3) \cdot \frac{1}{3} = -\frac{1}{3}$. Suppose that $F(u)$ has a ≥ 6-face.

If $\{k_1, k_2\} = \{3, 4\}$, then $t' \leq k-1-2t$, and so $\omega'(u) \geq \frac{k-3}{3} - \frac{1}{6} - \frac{1}{4} - \frac{1}{2} \cdot (t-1) - \frac{1}{3}(k-2-2t) = -\frac{1}{4} + \frac{t}{6} \geq -\frac{1}{12}$.

If $\min\{k_1, k_2\} = 3$ and $\max\{k_1, k_2\} \geq 6$, then $t' \leq k-2-2t$ unless $k \geq 5$ is odd and $t = \frac{k-1}{2}$. So $\omega'(u) = \frac{k-3}{3} - \frac{1}{4} - \frac{1}{2} \cdot (t-1) = \frac{k-6}{12} \geq -\frac{1}{12}$ if $k \geq 5$ is odd and $t = \frac{k-1}{2}$, and $\omega'(u) \geq \frac{k-3}{3} - \frac{1}{4} - \frac{1}{2} \cdot (t-1) - \frac{1}{3}(k-2-2t) = \frac{2t-1}{12} > 0$ otherwise.

If $k_1 = k_2 = 4$, then $t' \leq k-1-(2t+1) = k-2-2t$, $\omega'(u) \geq \frac{k-3}{3} - 2 \cdot \frac{1}{6} - \frac{1}{2} \cdot t - \frac{1}{3}(k-4-2t) = \frac{t}{6} \geq 0$.

If $k_1 = k_2 = 3$, then $t \geq 2$, $t' \leq k-2t$, and so $\omega'(u) \geq \frac{k-3}{3} - 2 \cdot \frac{1}{4} - \frac{1}{2} \cdot (t-2) - \frac{1}{3}(k-2t) = \frac{t-3}{6} \geq -\frac{1}{6}$.

Suppose $\max\{k_1, k_2\} \geq 6$ but $\min\{k_1, k_2\} \neq 3$. Then, $\omega'(u) \geq \frac{k-3}{3} - \frac{1}{2} \cdot t - \frac{1}{3}(k-2-2t) = \frac{t-2}{6} \geq -\frac{1}{6}$ if $t > 0$, $\omega'(u) \geq \frac{k-3}{3} - \frac{1}{6} - \frac{1}{3}(k-3) = -\frac{1}{6}$ if $t = 0$ and $\min\{k_1, k_2\} = 4$, and $\omega'(u) \geq \frac{k-3}{3} - \frac{1}{3}(k-3) = 0$ otherwise.

So, $\omega'(u) \geq -\frac{1}{6}$, and the equality may occur only in the last two cases where one face in $F(u) \cap N(f_o)$ is not a 4-face. ∎

Now, we proceed to estimate $\sum_{x \in V(G) \cup F(G)} \omega'(x)$. Let r_1 be the number of 3-vertices in C with $\omega' \geq -\frac{1}{4}$, r_2 be the number of ≥ 4-vertices in C with $\omega' \geq -\frac{1}{6}$, r_3 be the number of vertices in $V(G \setminus C)$ with $\omega' \geq \frac{1}{9}$, and let r_4 be the number of ≥ 8-faces.

If $\sum_{x \in V(G) \cup F(G)} \omega'(x) > -2$, we are done. Assume on the contrary that $\sum_{x \in V(G) \cup F(G)} \omega'(x) \leq -2$. Then by Claims 1, 2, and 3, $-2 \geq \frac{r-6}{6} - \frac{r_1}{4} - \frac{r_2}{6} - (r - r_1 - r_2) \cdot \frac{1}{3} + \frac{r_3}{9} + \frac{r_4}{3} = -1 - \frac{r}{6} + \frac{3r_1+6r_2+4r_3+12r_4}{36}$. So,

$$r = 7,\ r_4 = 0,\ \text{and } 3r_1 + 6r_2 + 4r_3 \leq 6. \tag{2}$$

We will show that C contains no ≥ 4-vertex, and then complete the proof by showing that $r_1 + r_3 \geq 2$ contradicts (2).

If $r_2 = 1$, suppose that $u \in C$ is the ≥ 4-vertex with $\omega'(u) \geq -\frac{1}{6}$. Then, $\omega'(u) = -\frac{1}{6}$, and $F(u) \cap N(f_o)$ has no 4-face. So, C has another ≥ 3-vertex incident with an inner ≥ 6-face that implies $r_1 \geq 1$ and contradicts (2).

Suppose that $r_2 = 0$. Then $\omega'(u) = -\frac{1}{3}$ for every ≥ 4-vertex in C. If C has a ≥ 4-vertex, say u, then $F(u)$ contains only 4-faces (by Claim 3). Suppose that $\{f_1, f_2\} = N(f_o) \cap F(u)$ (see Figure 14(a)). By Lemmas 5 and 7, for each $i \in \{1, 2\}$, $b(f_i) \setminus C$ has a ≥ 4-vertex, say u_i, which is incident with at most one 3-face of F_3' if it is a 5-vertex. So, $\omega'(u_i) \geq \frac{1}{9}$ (by Claim 2), $i = 1, 2$, and $r_3 \geq 2$, contradicting (2). Therefore, C has no ≥ 4-vertex.

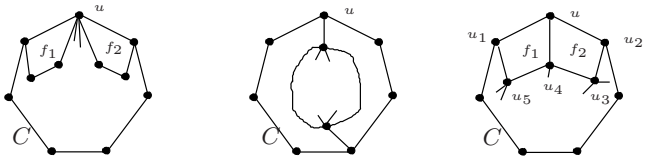


FIG. 14. $r = 7$, *a vertex* $u \in C$ *and its incident faces.*

If C has at most two 3-vertices (see Figure 14(b) as an example), then $r_4 \geq 1$ (since $r = 7$) and $r_1 = 2$ (by Claim 3), which contradicts (2). So, C has at least three 3-vertices. Since $r_1 \leq 2$ (by (2)), we may suppose that C has a 3-vertex, say u, such that $F(u)$ consists of two 4-faces f_1 and f_2 (see Figure 14(c)). By Lemma 7, $u_3, u_5 \notin C$.

Since $d(f_1) = d(f_2) = 4$, for each $i \in \{3, 4, 5\}$, if $d(u_i) = 5$, then it is incident with at most one 3-face of F_3', and hence $\omega'(u_i) \geq \frac{1}{9}$ if $d(u_i) \geq 4$ (by Claim 2). By Lemma 3, $d(u_4) = 3$ only if $d(u_3) \geq 4$ and $d(u_5) \geq 4$, so $d(u_4) \geq 4$ and $d(u_3) = d(u_5) = 3$ since $r_3 \leq 1$ by (2). Then, $r_3 = 1$.

If $F(u_1)$ (resp., $F(u_2)$) contains a ≥ 6-face, then $r_1 \geq 1$, contradicting (2). Otherwise, both $F(u_1)$ and $F(u_2)$ consist of two 4-faces, and the similar arguments as used for u show that $d(u_1) \geq 4$ and $d(u_2) \geq 4$, which together with u_4 give $r_3 \geq 3$. This contradicts (2) and totally completes the proof of Theorem 1. □

Acknowledgments. The author thanks the referees sincerely for their valuable suggestions, and thanks Dr. J. Huang for discussion about the examples as illustrated in Figure 1.

REFERENCES

[1] J. ANDREWS AND M. JACOBSON, *On a generalization of chromatic number*, Congr. Numer., 47 (1985), pp. 33–48.
[2] J. A. BONDY AND U. S. R. MURTY, *Graph Theory with Applications*, Elsevier, New York, 1976.
[3] O. V. BORODIN AND A. N. GLEBOV, *A sufficient condition for the* 3*-colorability of plane graphs*, Diskretn. Anal. Issled. Oper. Ser. 1, 11 (2004), pp. 13–29 (in Russian).
[4] O. V. BORODIN, A. N. GLEBOV, A. RASPAUD, AND M. R. SALAVATIPOUR, *Planar graphs without cycles of length from* 4 *to* 7 *are* 3*-colorable*, J. Combin. Theory Ser. B, 93 (2005), pp. 303–311.
[5] O. V. BORODIN AND A. RASPAUD, *A sufficient condition for planar graphs to be* 3*-colorable*, J. Combin. Theory Ser. B, 88 (2003), pp. 17–27.
[6] L. J. COWEN, R. H. COWEN, AND D. R. WOODALL, *Defective coloring of graphs in surfaces: Partitions into subgraphs of bounded valency*, J. Graph Theory, 10 (1986), pp. 187–195.
[7] L. COWEN, W. GODDARD, AND C. E. JESURUM, *Defective coloring revised*, J. Graph Theory, 24 (1997), pp. 205–219.
[8] M. FRICK, *A survey of* (m, k)*-coloring*, in Quo Vadis, Graph Theory?, J. Gimbel, J. W. Kennedy, and L. V. Quintas, eds., Ann. Discrete Math. 55, North–Holland, Amsterdam, 1993, pp. 45–57.
[9] F. HARARY AND K. JONES, *Conditional colorability II: Bipartite variations*, Congr. Numer., 50 (1985), pp. 205–218.
[10] T. JENSEN AND B. TOFT, *Graph Coloring Problems*, John Wiley and Sons, New York, 1995.
[11] R. STEINBERG, *The state of the three color problem*, in Quo Vadis, Graph Theory?, J. Gimbel, J. W. Kennedy, and L. V. Quintas, eds., Ann. Discrete Math. 55, North–Holland, Amsterdam, 1993, pp. 211–248.
[12] D. R. WOODALL, *List colourings of graphs*, in Surveys in Combinatorics, 2001, London Math. Soc. Lecture Note Ser. 288, Cambridge University Press, Cambridge, UK, 2001, pp. 269–301.

[13] B. Xu, *A* 3*-color theorem on plane graphs without* 5*-circuits*, Acta Math. Sin. (Eng. Ser.), 23 (2007), pp. 1059–1062.
[14] B. Xu, *On* 3*-colorable plane graphs without* 5*- and* 7*-cycles*, J. Combin. Theory Ser. B, 96 (2006), pp. 958–963.
[15] B. Xu and H. Zhang, *Every toroidal graph without adjacent triangles is* $(4,1)^*$*-choosable*, Discrete Appl. Math., 155 (2007), pp. 74–78.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 221–232

GEOMETRIC REALIZATION OF MÖBIUS TRIANGULATIONS*

MARÍA JOSE CHÁVEZ†, GAŠPER FIJAVŽ‡, ALBERTO MÁRQUEZ†, ATSUHIRO NAKAMOTO§, AND ESPERANZA SUÁREZ†

Abstract. A *Möbius triangulation* is a triangulation on the Möbius band. A *geometric realization* of a map M on a surface Σ is an embedding of Σ into a Euclidean 3-space $\mathbb{R}^3$ such that each face of M is a flat polygon. In this paper, we shall prove that every 5-connected triangulation on the Möbius band has a geometric realization. In order to prove it, we prove that if G is a 5-connected triangulation on the projective plane, then for any face f of G, the Möbius triangulation $G - f$ obtained from G by removing the interior of f has a geometric realization.

Key words. geometric realization, triangulation, Möbius band, projective plane

AMS subject classifications. 05C10, 52B70, 05C83

DOI. 10.1137/070693382

1. Introduction.

Let Σ be a surface with at most one boundary component, and let M be a map on Σ. If Σ has a boundary, we suppose that some cycle of M coincides with the boundary of Σ. Such a cycle of M is called the *boundary* of M and denoted by ∂M. A vertex of M not on ∂M is called an *inner* vertex. A *k-cycle* means a cycle of length k. A *triangulation* on Σ is a map on Σ such that each face is bounded by a 3-cycle. In particular, a *Möbius triangulation* is a triangulation on the Möbius band. For an inner vertex v of a triangulation, the *link* of v is the boundary walk of the 2-cell region consisting of all faces incident to v. Throughout this paper, we suppose that the graph of a map is *simple*, i.e., with no multiple edges and no loops. For a cycle or path C in M, a *chord* of C means an edge xy of M such that $x, y \in V(C)$ but $xy \notin E(C)$. Hence C is induced in M if and only if C has no chord.

A *geometric realization* of a map M on a surface Σ is an embedding of Σ into a Euclidean 3-space $\mathbb{R}^3$ such that each face of M is a flat polygon. Steinitz's theorem states that a spherical map has a geometric realization if and only if its graph is 3-connected [10]. Moreover, Archdeacon, Bonnington, and Ellis-Monanghan proved that every toroidal triangulation has a geometric realization [1]. In general, Grünbaum conjectured that every triangulation on any orientable closed surface has a geometric realization [7], but Bokowski and Guedes de Oliveira recently showed that a triangulation by K_{12} on the orientable closed surface of genus 6 has no geometric realization [2]. (For related topics, see [5].)

Let us consider a geometric realization of a triangulation on the projective plane. Let $\mathbb{P}$ denote the projective plane throughout this paper. Since the projective plane itself is not embeddable in $\mathbb{R}^3$, no map on $\mathbb{P}$ has a geometric realization. Let G be a triangulation on $\mathbb{P}$, and let f be a face of G. Let $G-f$ denote the Möbius triangulation

*Received by the editors May 31, 2007; accepted for publication (in revised form) August 22, 2008; published electronically December 19, 2008.
http://www.siam.org/journals/sidma/23-1/69338.html

†Departamento de Matematica Aplicada I, Universite de Sevilla, Escuela Universitaria Arquitectura Tecnica, Avda Reina Mercedes S/N, 41012 Sevilla, Spain (mjchavez@cica.es, almar@cica.es, emsuarez@cica.es).

‡Department of Computer Science, University of Ljubljana, 1000 Ljubljana, Slovenia (gasper.fijavz@fri.uni-lj.si).

§Department of Mathematics, Yokohama National University, Yokohama 240-8501, Japan (nakamoto@edhs.ynu.ac.jp).

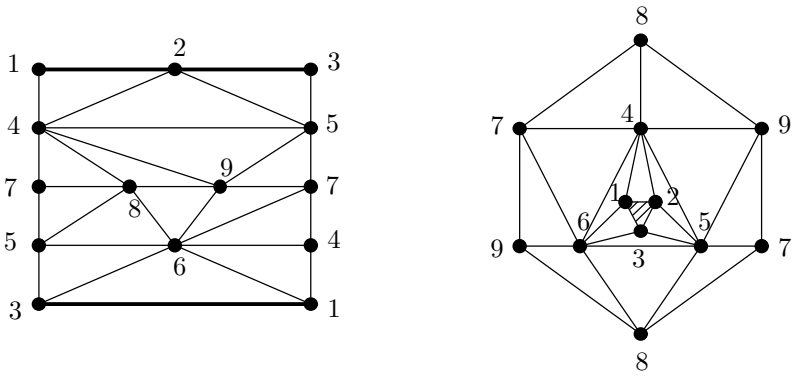


FIG. 1. *A Möbius triangulation with no geometric realization.*

obtained from G by removing the interior of f. Since the punctured surface obtained from $\mathbb{P}$ by removing a 2-cell, the *Möbius band*, is embeddable in $\mathbb{R}^3$, $G-f$ might have a geometric realization. The following is known.

THEOREM 1.1 (Bonnington and Nakamoto [3]). *Every triangulation G on the projective plane $\mathbb{P}$ has a face f such that the Möbius triangulation $G-f$ has a geometric realization.*

Brehm [4] has already found a Möbius triangulation with no geometric realization, shown in Figure 1, in which both express the same triangulation. (In Figure 1, we identify the vertices with the same label. In the right-hand side, the shaded part means the hole.) Why does Brehm's example have no geometric realization? We can prove that for each of its spatial embedding, the two disjoint 3-cycles 123 and 456 have a linking number of at least 2. (See [9] for the definition of the linking number.) However, two 3-cycles, each with an edge straight segment embedded in $\mathbb{R}^3$, have a linking number of at most 1, a contradiction. Hence, generalizing this example, we can see that if a triangulation M on the Möbius band has a boundary cycle C of length 3 and a 3-cycle C' disjoint from C which forms an annular region with C', then M never has a geometric realization.

A graph M is said to be *cyclically k-connected* if M has no separating set $S \subset V(M)$ with $|S| \leq k-1$ such that each connected component of $M-S$ has a cycle. Then the cyclical 4-connectivity of a triangulation G on $\mathbb{P}$ is necessary for a geometric realization of $G-f$ for any face f of G. We conjecture as follows that it is also sufficient.

CONJECTURE 1.2. *Let G be a triangulation on the projective plane $\mathbb{P}$. Then $G-f$ has a geometric realization for any face f of G if and only if G is cyclically 4-connected.*

In this paper, we prove the following.

THEOREM 1.3. *Let G be a 5-connected triangulation on the projective plane $\mathbb{P}$. Then $G-f$ has a geometric realization for any face f of G.*

By Theorem 1.3, a Möbius triangulation M has a geometric realization if M is obtained from a 5-connected triangulation G on $\mathbb{P}$ by removing a 2-cell.

Let M be a 5-connected Möbius triangulation with a boundary cycle $C = v_1 \cdots v_k$ of length k. Let P be the map on $\mathbb{P}$ obtained from M by pasting a 2-cell to C. If $k = 3$, then P is a 5-connected triangulation on $\mathbb{P}$. If $k = 4$, then P can be extended to a 5-connected triangulation on $\mathbb{P}$ by adding an edge v_1v_3 or v_2v_4. (If this is impossible, then M would have edges v_1v_3 and v_2v_4, and hence M would contain a quadrangulation

isomorphic to K_4, contrary to the 5-connectivity of M.) If $k \geq 5$, then P can be extended to a 5-connected triangulation on $\mathbb{P}$ by adding a new vertex joined to all vertices on C. Hence we have the following.

COROLLARY 1.4. *Every 5-connected Möbius triangulation has a geometric realization.*

Let M be a map on a surface Σ with a boundary, and let C be the boundary cycle of M. We say that M is *internally k-connected* if M is $(k-1)$-connected and if for any vertex $v \in V(M-C)$, there are at least k disjoint paths from v to C. Clearly, if G is a 5-connected triangulation on $\mathbb{P}$, then for any $v \in V(P)$, $G-v$ can be regarded as an internally 5-connected Möbius triangulation whose boundary cycle has a length of at least 5. Hence we can relax the condition of Corollary 1.4 to prove the following.

COROLLARY 1.5. *Every internally 5-connected Möbius triangulation has a geometric realization if the boundary cycle has a length of at least 5.*

2. Split-K_5's in 5-connected triangulations. Put a 5-cycle $C = v_1v_2v_3v_4v_5$ on $\mathbb{P}$, called the *boundary*, so that C bounds a 2-cell R on $\mathbb{P}$, where each v_i is called a *node*. (We always fix its orientation $\vec{C}$ along the numbering of the vertices.) Join v_i to v_{i+2} and v_{i+3} by edges not in R for each i. Then the resulting graph is isomorphic to K_5 in which each face except R is triangular. (See the left-hand side of Figure 2.) Consider a *splitting* (i.e., the inverse operation of an edge contraction) of v_i into two adjacent vertices, v_i and v_i', of degree 3. There are two possibilities for the splitting. When v_i and v_i' lie on C (we always suppose that v_i and v_i' appear on $\vec{C}$ in this order), $\{v_i, v_i'\}$ is called a *boundary pair* of nodes, and each of v_i and v_i' is called a *boundary split node*. (The path from v_i to v_i' on $\vec{C}$ is called the *boundary split interval* of $\{v_i, v_i'\}$.) Otherwise, $\{v_i, v_i'\}$ is called an *inner pair* of nodes, and each of v_i and v_i' is called an *inner split node*, where we always suppose that v_i lies on C. Let K be a map on $\mathbb{P}$ obtained from the above K_5 by splittings of some of v_i's. A *split-K_5* is a subdivision of K on $\mathbb{P}$. (See the right-hand side of Figure 2.)

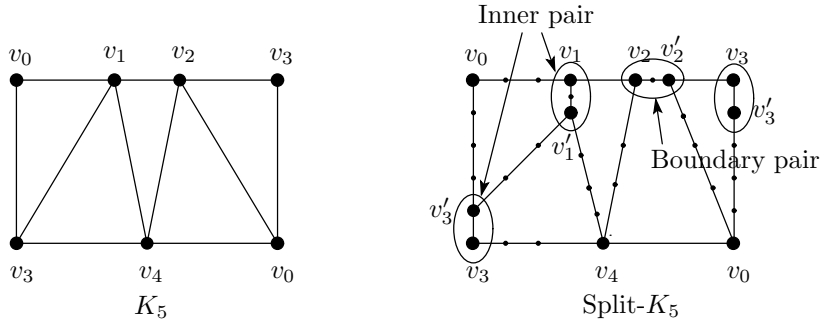


FIG. 2. *K_5 and split-K_5.*

The following is the most important claim in this paper. It guarantees that a 5-connected triangulation on $\mathbb{P}$ has a special type of a split-K_5.

LEMMA 2.1. *Let G be a 5-connected triangulation on $\mathbb{P}$, and let uvw be any face of G. Then G has a split-K_5 H such that*

(i) *the boundary ∂H of H coincides with the link of u in G.*
(ii) *H has at most one boundary pair of nodes.*
(iii) *if H has a boundary pair, then at least one of v and w is a boundary split node, but the edge vw is not contained in a boundary split interval. Otherwise, v or w is a node of H.*

In the following two sections, we give preliminaries for the proof of Lemma 2.1. In section 5, we prove Lemma 2.1.

3. Lemmas. Let G be a graph on $\mathbb{P}$, and let C be a *contractible* cycle of G, i.e., one bounding a 2-cell on $\mathbb{P}$. (A cycle or a closed curve on a surface is *essential* if it is not contractible.) Then C cuts $\mathbb{P}$ into two surfaces, one homeomorphic to an open disk and the other homeomorphic to an open Möbius band. Let $\mathrm{int}_C(G)$ denote the graph consisting of the vertices and edges lying in the disk component of C, and let $\mathrm{Int}_C(G)$ be the graph consisting of the vertices and edges lying on C and in the disk component of C. We define $\mathrm{ext}_C(G)$ and $\mathrm{Ext}_C(G)$ analogously. Note that $\mathrm{Int}_C(G)$ is not necessarily an induced subgraph of G.

Let $C = v_1v_2v_3v_4\cdots v_k$ be a cycle. A *closed segment* $[v_i, v_j]$ is a $v_i - v_j$ path along $\vec{C}$. An *open segment* (v_i, v_j) is obtained by deleting the endvertices of the corresponding closed segment. Moreover, we use the notations $[v_i, v_j)$ and $(v_i, v_j]$, defined similarly.

LEMMA 3.1. *Let G be a 5-connected triangulation on $\mathbb{P}$. Let $C = v_1v_2v_3v_4$ be a contractible 4-cycle in G. Then $\mathrm{int}_C(G)$ contains no vertices.*

Proof. Assume $v \in V(\mathrm{int}_C(G))$. Since G is 5-connected, $\mathrm{ext}_C(G)$ contains no vertices. Then we can add only two edges v_1v_3 and v_2v_4 outside C, since T is simple. Hence this contradicts that G is a triangulation. ∎

LEMMA 3.2. *Let v be a vertex of a 5-connected triangulation G on $\mathbb{P}$, and let C be a contractible 5-cycle containing v in its interior. Then there exists a unique contractible 5-cycle $\overline{C}$ so that $\mathrm{Int}_{\overline{C}}(G)$ contains all contractible 5-cycles which contain v in their respective interiors.*

Proof. Let C_1 and C_2 be contractible 5-cycles containing v in their interiors, and suppose that $\mathrm{Int}_{C_1}(G)$ and $\mathrm{Int}_{C_2}(G)$ are *inclusionwise incomparable*, that is, neither $\mathrm{Int}_{C_1}(G) \subseteq \mathrm{Int}_{C_2}(G)$ nor $\mathrm{Int}_{C_1}(G) \supseteq \mathrm{Int}_{C_2}(G)$. It suffices to prove that there is a contractible 5-cycle C' such that $\mathrm{Int}_{C'}(G)$ contains both $\mathrm{Int}_{C_1}(G)$ and $\mathrm{Int}_{C_2}(G)$.

Since C_1 and C_2 are of length 5 and neither one is contained in the closed interior of the other, they intersect in exactly two vertices. These two vertices divide C_i into a segment lying in the interior of C_{3-i} and one lying in the exterior of C_{3-i}, where $i = 1, 2$. Combining the common segments and both interior segments yields a contractible cycle, which contains v in its interior. By Lemma 3.1, its length is at least 5. Combining the two exterior segments with the two common segments, we obtain a contractible cycle C' of length at most 5, since both C_1 and C_2 were 5-cycles. Since G is simple, C' contains no essential cycle, and hence it is a contractible cycle in G. Now C' has length 5 by Lemma 3.1 since it contains v in its interior. On the other hand, $\mathrm{Int}_{C'}(G)$ contains both $\mathrm{Int}_{C_1}(G)$ and $\mathrm{Int}_{C_2}(G)$, and the proof is complete. ∎

LEMMA 3.3. *Let G be a 5-connected triangulation on $\mathbb{P}$, and let $C = v_1v_2v_3v_4v_5$ be a contractible 5-cycle in G. If G has no vertex in the exterior of C, then $\mathrm{Ext}_C(G)$ is isomorphic to K_5.*

Proof. We have to show that $\mathrm{ext}_C(G)$ contains every possible edge v_iv_{i+2} (in indices modulo 5). A similar argument as in the proof of Lemma 3.1 does the trick. ∎

The following lemma is an immediate consequence of 5-connectivity.

LEMMA 3.4. *Let G be a 5-connected triangulation on $\mathbb{P}$, and let $v \in V(G)$. Let v' and v'' be two nonconsecutive neighbors of v. If v' and v'' have another common neighbor w which is not adjacent to v, then the cycle $vv'wv''$ is essential.*

Let D be a plane graph with boundary cycle C and each inner face triangular, and let x, y be distinct vertices of C. An *internal $x - y$ path* is a path in D joining x and y and intersecting C only at its endvertices.

LEMMA 3.5. *Let D be a triangulation on the disk with boundary cycle C, and let x, y be distinct vertices of C with $xy \notin E(C)$. Then D has an internal $x-y$ path if and only if D has no chord pq for some $p, q \in V(D) - \{x, y\}$ such that x and y are contained in distinct components of $C - \{p, q\}$.*

Proof. The sufficiency is obvious and so we consider the necessity. Suppose that C has a chord pq. By the assumption, x and y are contained in one, say D_1, of the two subgraphs D_1, D_2 such that $V(D) = V(D_1) \cup V(D_2)$ and $V(D_1) \cap V(D_2) = \{p, q\}$. In this case, we have to look for a required internal $x-y$ path in D_1. Hence in the following argument, we may suppose that D has no chord. Observe that since C is chordless, each vertex on C is adjacent to at least one vertex in $D - C$. Moreover, we can see that $\mathrm{int}_C(D)$ is connected. (For otherwise, i.e., if $\mathrm{int}_C(D)$ is disconnected, then there are two vertices $p', q' \in C$ such that $D - \{p', q'\}$ is disconnected. However, this is impossible since each inner face of D is triangular.) Hence we have an internal $x-y$ path in D. □

Let C be a contractible cycle of length at least 4 in a triangulation G. Suppose that vertices r_1, r_2, r_3, r_4 lie along C in this order, but they do not need to be consecutive along C. Let us also assume that the segments $[r_1, r_2]$, $[r_2, r_3]$, $[r_3, r_4]$, and $[r_4, r_1]$ have no chords in $\mathrm{Int}_C(G)$. We say that $\mathrm{Int}_C(G)$ is a 4-*patch with nodes* r_1, r_2, r_3, r_4.

We obtain the following three lemmas, carefully applying Lemma 3.5 to P.

LEMMA 3.6. *Let P be a 4-patch with nodes r_1, r_2, r_3, r_4. Assume that $r_1r_4, r_2r_3 \in E(P)$ and that u and v are vertices from (r_1, r_2) and (r_3, r_4), respectively. Then $P - \{r_1, r_2, r_3, r_4\}$ contains an $u-v$ path, or a pair of antipodal nodes are adjacent.*

LEMMA 3.7. *Let P be a 4-patch with nodes r_1, r_2, r_3, r_4. Then $P - \{r_1, r_3\}$ contains an $r_2 - r_4$ path unless $r_1r_3 \in E(P)$.*

Let P be a 4-patch with nodes r_1, r_2, r_3, r_4. An $r_2 - r_4$ *diagonal* in P is an $r_2 - r_4$ path $Q = u_1u_2u_3 \cdots u_{k-1}u_k$ ($u_1 = r_2$ and $u_k = r_4$) in $P - \{r_1, r_3\}$ if there exists indices $i < j$ such that

(D1) the *initial segment* $u_1 \cdots u_i$ is a segment of ∂P,
(D2) the *terminal segment* $u_j \cdots u_k$ is a segment of ∂P, and
(D3) the *intermediate segment* $u_i \cdots u_j$ is a segment of P such that $u_i, u_j \in V(\partial P)$ and that all other vertices lie in $\mathrm{int}(P)$.

If Q is an $r_2 - r_4$ diagonal in P, then it is also an $r_4 - r_2$ diagonal. Further, if a patch P with nodes r_1, r_2, r_3, r_4 contains an $r_2 - r_4$ path avoiding r_1 and r_3, then it also contains an $r_2 - r_4$ diagonal.

We say that an $r_2 - r_4$ diagonal Q lies *closest to* r_1 if the number of faces of P bounded by Q and the segments incident with r_1 is as small as possible.

LEMMA 3.8. *Let P be a 4-patch with nodes r_1, r_2, r_3, r_4, and let Q be the $r_2 - r_4$ diagonal closest to r_1. Let u_i and u_j be the first and last vertex of the intermediate segment of Q, respectively. Then r_1 is adjacent to $u_i, u_{i+1}, \ldots, u_{j-1}, u_j$ in P.*

4. Essential 3-linkages. A *near triangulation* R is a map on $\mathbb{P}$ with a distinguished face f such that every other face of R is triangular, and that the facial walk along f is a cycle. Suppose that the boundary cycle of f, denoted by W, has a length of at least 6. Let $v_1, v_2, v_3, v_4, v_5, v_6$ be six vertices that appear along W in this order but that do not need to be consecutive along W. An *essential* 3*-linkage* (with respect to $v_1, v_2, v_3, v_4, v_5, v_6$) is a collection L of three disjoint paths P_1, P_2, P_3 so that P_i is a $v_i - v_{i+3}$ path for $i = 1, 2, 3$. It is easy to see that $W \cup P_i$ contains some essential cycle. Let Q_1 be some minimal subpath of P_1 so that $W \cup Q_1$ still contains an essential cycle. Also Q_1, P_2, P_3 form an essential 3-linkage with possibly different endvertices. By applying the same idea on P_2 and P_3, we obtain the following lemma.

LEMMA 4.1. *Let L be an essential 3-linkage with respect to nodes $v_1, \dots, v_6$. There exists an essential 3-linkage L' so that every path in L' intersects W only at its endvertices.*

The second result has been, in greater generality, proved by Robertson and Seymour in [8]. We state it adapted to our needs.

THEOREM 4.2 (Robertson and Seymour [8]). *Let R be a near triangulation of $\mathbb{P}$ and $f = v_1v_2v_3v_4v_5v_6$ its distinguished face of length 6. Then R contains an essential 3-linkage with respect to $v_1, v_2, v_3, v_4, v_5, v_6$ if and only if*

(L1) *R contains no pair of parallel nonhomotopic edges with common endvertices;*
(L2) *R does not contain a contractible cycle C of length at most 5 whose interior contains f.*

A pair of parallel nonhomotopic edges violating (L1) forms an essential cycle of length 2. Traversing these two edges twice yields a contractible (but not simple) closed walk whose "interior" contains all faces of R. This observation enables both conditions (L1) and (L2) to be combined into a single condition, albeit with slight adaptations. For practicality, we prefer the conditions to be written separately, since they are of different flavors and have to be tackled with different approaches.

We look for essential 3-linkages in near triangulations. In the case when the length of the distinguished face exceeds 6, we first decide which six vertices are the endvertices of a linkage. The rest of this section is devoted to the proof of the following.

PROPOSITION 4.3. *Let G be a 5-connected triangulation of $\mathbb{P}$, and let v be a vertex of degree $d \geq 6$. Let $D = u_1u_2 \cdots u_d$ be the link of v in G. Then the near triangulation $R = G - v$ contains an essential 3-linkage if and only if v is not contained in the interior of a contractible cycle of length at most 5.*

Proof. Clearly a cycle containing v in its interior meets each path in an essential 3-linkage at least twice. The difficulty lies in the other direction—how to find a linkage—if v is not contained in the interior of a "short" contractible cycle.

An edge $e \in E(R)$ is said to be *essential* if the endvertices of e lie in D and $D \cup e$ contains an essential cycle. We shall split the proof of Proposition 4.3 with respect to the number of essential edges. If R contains a set of three independent essential edges, then no further proof is needed. This leaves us with the case where a maximal set of independent essential edges contains at most two edges.

Assume next that R contains a set of two independent essential edges. The four endvertices of these essential edges split the f-facial walk into four open segments. Let us choose essential edges $e = r_1r_4$ and $e' = r_3r_6$ in such a way that the union of two consecutive open segments $(r_1, r_6) \cup (r_3, r_4)$ in D contains as few vertices as possible. Suppose that (r_1, r_3) contains a vertex, say v_2, and that (r_4, r_6) contains a vertex, say v_5. Now if $r_1r_6 \in E(R) - E(D)$, then the contractible cycle $vr_4r_1r_6$ separates v_2 from v_5, and if $r_3r_4 \in E(R) - E(D)$, then the contractible cycle $vr_3r_4r_1$ separates v_2 from v_5. Neither can happen since G is 5-connected. By Lemma 3.6, we can join v_2 and v_5 by a path avoiding r_1, r_2, r_3, and r_4, and hence we can find an essential 3-linkage.

So we assume that there exists a set of two independent essential edges $e = w_1w_3$ and $e' = w_2w_4$ so that w_1, w_2, and w_3 lie consecutively along D. We may also assume that w_4 lies closer to w_3 than to w_1 along D, and that no essential edge incident with w_2 has the other endvertex in (w_3, w_4). Denote the vertices along D by $v_1, v_2, v_3, \dots, v_d$ so that $v_1 = w_1$ and $v_2 = w_2$ (also $v_3 = w_3$, but then this may not go on). Add to R the *new* edges v_1v_k, where $k = 6, \dots, d-1$, and denote the resulting near triangulation with R', with the distinguished face of size 6.

It is easy to see that R' satisfies (L1), since the newly added edges do not have their essential counterparts. Similarly, a short contractible cycle C containing the distinguished face of R' in its interior, i.e., contradicting (L2), would have to use some new edge v_1v_k, where $k \geq 6$. Now C would contain vertices v_k, w_1, w_2, and w_3, which implies that vertices v_k and w_3 have a common neighbor in R. This contradicts Lemma 3.4 since C is contractible. Hence R' contains an essential 3-linkage. Since all new edges share a common endvertex, we can, if necessary, transform the linkage into an essential 3-linkage in R.

Suppose next that there is an essential edge but we cannot find a set of two independent essential edges. Let $e = w_1w_2$ be the essential edge, and assume that the segment (w_1, w_2) is as short as possible. Since G is simple, w_1 and w_2 are not consecutive along D. Denote the vertices of D so that $w_1 = v_3$ and v_4 lies in (w_1, w_2). As (w_1, w_2) is as short as possible, we have $w_2 \neq v_1$.

As in the previous case, let R' be the near triangulation obtained by adding *new* edges v_1v_k, where $k = 6, \ldots, d-1$. We will argue that R' has an essential 3-linkage.

If R' does not satisfy (L1), then an essential edge e' must be incident with both v_1 and v_k for some k satisfying $6 \leq k \leq d$. By interlacing essential edges incident to $v_k \in [w_1, w_2] = [v_3, w_2]$, we clearly have $v_k \neq v_3$. On the other hand, v_k cannot lie in $(w_1, w_2) = (v_3, w_2)$, as two independent essential edges cannot exist, and hence $v_k = w_2$. But this contradicts 5-connectivity of G, since the 4-cycle $vv_1v_kv_3 = vv_1w_2w_1$ separates v_2 and v_4.

Next assume that R' contradicts (L2). The short cycle C contradicting (L2) can be divided into three segments: the first one between v_1 and w_1, the second between w_1 and w_2, and the third between w_2 and v_1. Their lengths are at least 2, 2, and 1, respectively, using the fact that neither v_1 and $w_1 = v_3$ nor w_1 and w_2 are consecutive along D, and the fact that C uses one of the new edges. Since the length of C is at most 5, all lower bounds are sharp. By Lemma 3.4, C must pass through v_2, and also C must pass through v_4 and $w_2 = v_5$. On the segment between w_2 and v_1 the cycle C uses exactly one edge, namely $v_1w_2 = v_1v_5$, and it also has to use one new edge. This is a contradiction, so R' satisfies both (L1) and (L2), and R' contains an essential 3-linkage. As in the previous case we can, if necessary, transform the linkage into an essential 3-linkage in R.

We are left with the case where R contains no essential edges. Even if we add new edges to the interior of f, we cannot contradict (L1), and our only concern will be meeting the condition (L2).

We proceed naively. Let us assign labels $v_1, v_2, \ldots, v_d$ to neighbors of v in the order of their indices. Add new edges of the form v_1v_k, where $k = 6, \ldots, d-1$. The newly obtained near triangulation R' may contain an essential 3-linkage, and we win. On the other hand, it may not, as we contradict (L2), and we *lose*. In this case, R' contains a short cycle C which uses a new edge v_1v_ℓ for some $\ell \in \{6, \ldots, d-1\}$.

Hence we assume that we lose for every assignment of labels $v_1, v_2, \ldots, v_d$ to the consecutive neighbors of v. Now fix an assignment of labels so that there exists a cycle C_v contradicting (L2) using a new edge v_1v_k, where k is as large as possible.

Let us denote $w_1 = v_1$, $w_2 = v_2$, $w_3 = v_3$, $w_4 = v_{k-1}$, $w_5 = v_k$, and $w_6 = v_{k+1}$. Further, let us add new edges joining w_1 to vertices of (w_6, w_1) and additional new edges joining w_3 to vertices of (w_3, w_4). We denote the newly obtained near triangulation by R_w. We claim that R_w contains an essential 3-linkage.

Assume that this is not the case, and let C_w be the obstruction according to (L2). Clearly C_w contains at least one new edge. Observe that C_w cannot contain both a

new edge incident with w_1 and a new edge incident with w_3, since a segment of C_w of length at most 2 would join two nonconsecutive vertices of D. The cycle C_w cannot contain a new edge incident with w_1 since this would contradict maximality of k. Hence, C_w contains a new edge incident with w_3. Now let C' be the cycle containing the edges of C_w lying outside C_v and the edges of C_v lying outside of C_w. Then C' is a contractible cycle containing f in its interior. Let $P \subseteq C_w \cup C_v$ be the $v_1 - v_3$ path whose edges lie in the interior of C'. Since it connects two nonconsecutive vertices along f, its length is at least 3. This implies that the length of C' is at most 5, a contradiction.

Hence R_w contains an essential 3-linkage, and consequently R also contains an essential 3-linkage. This completes the proof of Proposition 4.3. □

5. Proof of Lemma 2.1. In this section, we shall prove Lemma 2.1. We begin with the following proposition.

PROPOSITION 5.1. *Let G be a 5-connected triangulation on $\mathbb{P}$, and let $u \in V(G)$. Then G has a split-K_5 H whose boundary coincides with the link of u in G.*

Proof. We will split the analysis into two cases regarding the properties of u and treat one of the two cases by referring to [6]. Let D be the link of u.

Case 1. G contains a contractible 5-cycle $C = v_1v_2v_3v_4v_5$ such that $u \in V(\mathrm{int}_C(G))$.

By Lemma 3.2, we may assume that C is the maximal 5-cycle containing u in its interior. Since G is 5-connected, there exist internally disjoint $u - v_i$ paths P_i for $i = 1, \ldots, 5$.

In order to find a suitable split-K_5, we need to find a subgraph of $\mathrm{Ext}_C(G)$ which contracts to the zigzag cycle $v_1v_3v_5v_2v_4$. This task has been treated in greater generality in [6, subsection: Finding a suitable cycle minor U in G_x]. Hence we can obtain a split-K_5 H' whose boundary is C. Now let

$$H = (H' - E(C)) \cup D \cup \bigcup_{i=1}^{5} (P_i - \{v\}).$$

Then H is a split-K_5 with boundary D, in which there is no boundary pair.

Case 2. u does not lie in the interior of a contractible 5-cycle.

Then we clearly have $|D| = \deg(u) = k \geq 6$. Let f be the distinguished face of $G - v$ with boundary D. By Theorem 4.2, $G - v$ contains an essential 3-linkage $L = \{P_1, P_2, P_3\}$ with respect to $u_1, u_2, u_3, u_4, u_5, u_6$, where P_i joins u_i and u_{i+3} for $i = 1, 2, 3$. We may also assume that each P_i in L has no chord. Then L divides the near triangulation $G - v$ into three patches R_{12}, R_{23}, and R_{13}, whose nodes are (u_1, u_2, u_5, u_4), (u_2, u_3, u_6, u_5), and (u_3, u_4, u_1, u_6) lying on their boundary in this order, respectively.

We first claim that these patches contain two vertex-disjoint diagonals. Let us first prove that every two patches, say R_{12} and R_{23}, contain diagonals with disjoint endvertices. Suppose this is not the case, and let, say, u_2 be an endvertex of every possible diagonal in both R_{12} and R_{23}. By Lemma 3.7, we have $u_2u_4 \in E(R_{12})$ and $u_2u_6 \in E(R_{23})$. This contradicts the 5-connectivity of G since $\{u, u_2, u_4, u_6\}$ separates v_5 and v_1 in G. Hence we may assume that R_{12} contains a $u_1 - u_5$ diagonal D_{15} and that R_{23} contains a $u_2 - u_6$ diagonal D_{26}. We first suppose that D_{15} and D_{26} are disjoint. In this case, we can obtain a required split-K_5 H such that $H = D \cup L \cup D_{15} \cup D_{26}$.

Now consider the case when D_{15} and D_{26} share an inner vertex. Let us try to push the diagonals away: suppose that D_{15} and D_{26} are closest to u_4 and u_3, respectively. If D_{15} and D_{26} are not vertex disjoint, then the terminal segment S of D_{15} intersects

the initial segment S' of D_{26} at P_2. Let w be the first vertex of S, and let w' be the last vertex of S'. Then, by Lemma 3.8, we have both $u_4w \in E(R_{12})$ and $u_3w' \in E(R_{23})$. If $w \neq w'$, then we can find a $u_2 - u_4$ diagonal in R_{12} through wu_4 and a $u_3 - u_5$ diagonal in R_{23} through u_3w'. Since they are disjoint, we are done, similarly as above.

Suppose that $w = w'$. Since $u_4w \in E(R_{12})$, we focus on the 4-patch R'_{12} with nodes u_1, u_2, w, u_4 contained in R_{12}. Note that $u_1w \notin E(R'_{12})$. (For otherwise, $\{u, u_1, w, u_3\}$ separates u_2 and u_4, since $u_3w \in E(R_{23})$. This contradicts the 5-connectivity of G.) Hence R'_{12} admits a $u_2 - u_4$ diagonal D_{24}, avoiding w and u_1, by Lemma 3.7. Let D_{35} be the $u_3 - u_5$ diagonal of R_{23} through u_3w. Then $D \cup L \cup D_{24} \cup D_{35}$ is a required split-K_5 in G since D_{24} and D_{35} are disjoint. □

By Proposition 5.1, a 5-connected triangulation on $\mathbb{P}$ has a split-K_5 H whose boundary coincides with the link of a specified vertex. Let $[a, b]$ denote the path in H joining two vertices a and b which is contained in the path joining two nodes in H, where $1 \leq i < j \leq 5$. Moreover, we denote $(a, b) = [a, b] - \{a, b\}$, and also use the notations $[a, b)$ and $(a, b]$ similarly.

The following claims that a boundary pair of nodes can be "moved" in a sense.

LEMMA 5.2. *Suppose that a triangulation G on $\mathbb{P}$ has a split-K_5 H with boundary C. Let $\{a', a''\}$ be a boundary pair of nodes of H, and let Q be the plane subgraph of G corresponding to a face of H with nodes a', a'', b, c. Then, for some vertex a of $[a', a'']$ in G, we can find a split-K_5 H' with boundary C such that a is a node of H' contained in neither a boundary pair nor an inner pair. Moreover, if b is contained in a boundary pair, then the number of the boundary pairs can be decreased in H'; otherwise, b might be contained in a new boundary pair of H'.*

Proof. We may suppose that a vertex y of $(a', c]$ and a vertex z of $(a', a'']$ are not adjacent in Q. (For otherwise, replacing $[a', y)$ with zy, we can regard z as a new a'.) Then, by Lemma 3.5, we can take an internal $a' - x$ path P for some x on either $(a'', b]$ or (b, c). In the former case, let $H' = H - (a'', x) \cup P$ (or $H' = H - (a'', b') \cup P$ when x is in (b, b') for an inner pair $\{b, b'\}$). See Figure 3. Then we can decrease the number of boundary split pairs. In the latter case, let $H' = H - (a'', b) \cup P$ (or $H' = K - (a'', b') \cup P$ when $\{b, b'\}$ is an inner pair), in which x might be a new boundary pair. □

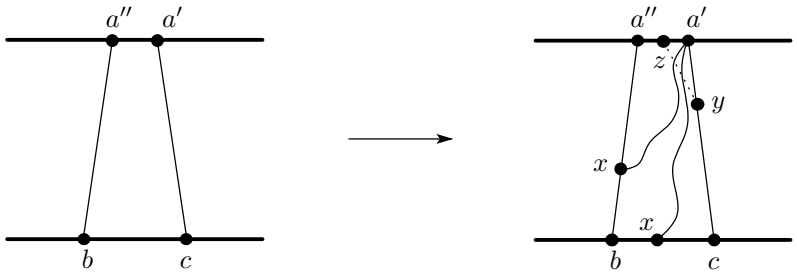


FIG. 3. *Eliminate or move a boundary split node.*

Now we shall prove Lemma 2.1.

Proof of Lemma 2.1. Let G be a 5-connected triangulation on $\mathbb{P}$, and let uvw be any face of G. By Proposition 5.1, G has a split-K_5 H whose boundary ∂H coincides with the link of u in G. Let v_1, v_2, v_3, v_4, v_5 be five nodes of H (where $\partial\vec{H}$ is fixed along the ordering of $v_1, \ldots, v_5$); some v_i's might be contained in boundary or inner pairs $\{v_i, v'_i\}$ of nodes.

We shall deform H to satisfy conditions (ii) and (iii) in the lemma. We may

suppose that the edge vw is contained in $[v_1, v_2]$ so that $v\vec{w}$ is along $\partial\vec{H}$. Moreover, we may suppose that neither v_1 nor v_2 is a boundary split node. (For otherwise, we can apply Lemma 5.2 to $\{v_1, v_1'\}$ or $\{v_2, v_2'\}$.)

We first show that one of v and w can be chosen as a node in a new split-K_5. Hence we may suppose that $v \neq v_1$ and $w \neq v_2$. Let R be the plane subgraph of G corresponding to a face of H incident to $[v_1, v_2]$. Suppose that R is bounded by $[v_1, v_2]$, $[v_1, v_4]$, $[v_4, v_4']$, and $[v_2, v_4']$ of H, when $\{v_4, v_4'\}$ is a boundary split pair. (See Figure 4. Since the other two cases shown in the figure are similar, we omit the details.) Observe that there are no two vertices x and y in $[v_1, v_2]$ joined by a chord. (For otherwise, $\{x, y, u\}$ would be a 3-cut of G, contrary to the 5-connectivity of G.) Hence, by Lemma 3.5, we can find an internal path P from v to a vertex on $(v_1, v_4]$, to a vertex on $(v_4, v_4']$, or to $(v_2, v_4']$. In the first and second cases, adding P to H and deleting a segment suitably, we obtain a split-K_5 with v a node. If we do not have these cases, then there is a vertex s in $[v_1, v)$ and a vertex t in (v_4', v_2) which are adjacent in R. In this case, we must have an internal path P' from w to some vertex r of (v_4', v_2) in R. Similarly to the previous two cases, we obtain a split-K_5 with w a node.

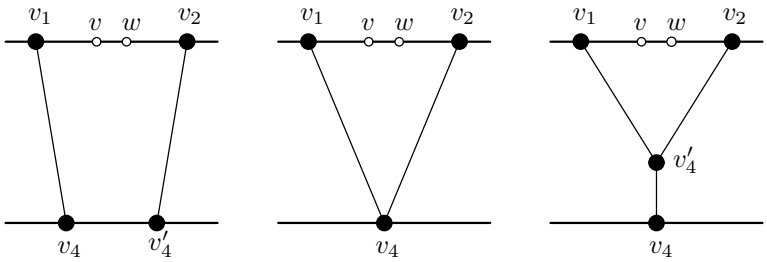


FIG. 4. *Take a path from v or w.*

We may suppose that v is a node. If v is a boundary split node, then put $v = v_1'$, and suppose that vw is contained in $[v_1', v_2]$. Otherwise, put $v = v_1$. If v_4 is contained in a boundary pair $\{v_4, v_4'\}$, then we apply Lemma 5.2 to eliminate the boundary pair $\{v_4, v_4'\}$, fixing v, or move the boundary pair toward v_2. (Otherwise, we proceed to v_2.) Then, fixing the new v_4, we apply Lemma 5.2 to $\{v_2, v_2'\}$ if $\{v_2, v_2'\}$ is a boundary split pair. Similarly, we apply Lemma 5.2 to $\{v_5, v_5'\}$ and $\{v_3, v_3'\}$ in this order if necessary. Then, the resulting split-K_5 has at most one boundary split pair containing v.

6. Proof of the theorem. In this section, we shall prove Theorem 1.3. The main part of the proof, which is to make a geometric realization of a 5-connected triangulation G on $\mathbb{P}$ with any one face f removed, depends on the technique developed in [3].

LEMMA 6.1 (Bonnington and Nakamoto [3]). *Let T be a Möbius triangulation with boundary C. Suppose that T has a split-K_5 H with boundary C and at most one boundary pair of nodes.*

(i) *If H has no boundary pair and we let v_1, v_2, v_3, v_4, v_5 be the nodes of T lying on C in this order, then let e be the edge of $[v_1, v_2]$ incident to v_1.*

(ii) *If H has a boundary pair $\{v_1, v_1'\}$ and we let $v_1, v_1', v_2, v_3, v_4, v_5$ be the nodes of T lying on C in this order, then let e be the edge of $[v_1', v_2]$ incident to v_1'.*

Then T has a geometric realization $\hat{T}$ such that all edges on C except e can be seen from some fixed point $x \in \mathbb{R}^3$.

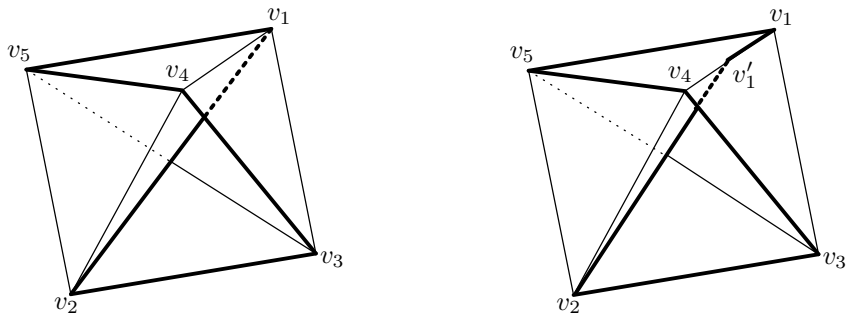


FIG. 5. *Examples of geometric realizations of* T.

Figure 5 shows examples of geometric realizations of split-K_5's satisfying Lemma 6.1. The left-hand side shows one with exactly five nodes v_1, v_2, v_3, v_4, v_5 on the boundary, and the right-hand side shows one with exactly one boundary split pair $\{v_1, v_1'\}$. (Note that a triangulation G dealt with in Lemma 6.1 might have several inner pairs of nodes.) In both parts of figure, we can see all segments on ∂H, except a side of $[v_1, v_2]$ incident to v_1 in the left-hand case and a side of $[v_1', v_2]$ incident to v_1' in the right-hand case.

Now we shall prove Theorem 1.3.

Proof of Theorem 1.3. Let G be a 5-connected triangulation on $\mathbb{P}$, and let f be any face of G bounded by uvw. Let C be the link of u. Then, by Lemma 2.1, G contains a split-K_5 H such that

(i) the boundary ∂H of H coincides with C,
(ii) H has at most one boundary split pair, and
(iii) if H has a boundary pair, then v is a boundary split node of H, but vw is not contained in a boundary split interval; otherwise, v or w is a node of H.

Consider the Möbius triangulation $G' = G - u$ with boundary C. We apply Lemma 6.1 to G' and the above H. Then we get a geometric realization $\hat{G}'$ of G' such that from some point $x \in \mathbb{R}^3$, all edges on C except vw can be seen.

First, we put the vertex u at $x \in \mathbb{R}^3$. For each edge pq of $\hat{G}'$ lying on C, let $\Delta_{pq} \in \mathbb{R}^3$ denote the triangular disk with x, p, q as its vertices. Now, for any edge $h \in E(C) - \{vw\}$, we shall fit Δ_h into the body of $\hat{G}'$, where Δ_h corresponds to a face of G incident to h and v. Since each $h \in E(C) - \{vw\}$ can be seen from $x \in \mathbb{R}^3$, the interior of Δ_h does not collide with $\hat{G}'$. Moreover, for any two distinct $h, h' \in E(C) - \{vw\}$, the interiors of Δ_h and $\Delta_{h'}$ do not collide internally, since h and h' can be seen from x simultaneously. So we get a geometric realization of $G - f$. ∎

Acknowledgments. The authors are grateful to two anonymous referees for their carefully reading of the paper and helpful suggestions.

REFERENCES

[1] D. ARCHDEACON, C. P. BONNINGTON, AND J. A. ELLIS-MONANGHAN, *How to exhibit toroidal maps in space*, Discrete Comput. Geom., 38 (2007), pp. 573–594.
[2] J. BOKOWSKI AND A. GUEDES DE OLIVEIRA, *On the generation of oriented matroids*, Discrete Comput. Geom., 24 (2004), pp. 197–208.
[3] C. P. BONNINGTON AND A. NAKAMOTO, *Geometric realization of a triangulation on the projective plane with one face removed*, Discrete Comput. Geom., 40 (2008), pp. 141–157.

[4] U. Brehm, *A nonpolyhedral triangulated Möbius strip*, Proc. Amer. Math. Soc., 89 (1983), pp. 519–522.
[5] U. Brehm and J. M. Wills, *Polyhedral manifolds*, in Handbook of Convex Geometry, P. M. Gruber and J. M. Wills, eds., North-Holland, Amsterdam, 1993, pp. 535–554.
[6] G. Fijavž and B. Mohar, *K_6-minors in projective planar graphs*, Combinatorica, 23 (2003), pp. 453–465.
[7] B. Grünbaum, *Convex Polytopes*, Pure and Appl. Math. 16, Interscience Publishers John Wiley & Sons, Inc., New York, 1967.
[8] N. Robertson and P. D. Seymour, *Graph minors.* VI. *Disjoint paths across a disc*, J. Combin. Theory Ser. B, 41 (1986), pp. 115–138.
[9] D. Rolfsen, *Knots and Links*, Math. Lecture Ser. 7, Publish or Perish, Inc., Berkeley, CA, 1976.
[10] E. Steinitz, *Polyeder und Raumeinteilunger*, Enzykl. Math. Wiss., 3 (1922), part 3A612, pp. 1–139.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 233–250

CLEANING REGULAR GRAPHS WITH BRUSHES*

NOGA ALON†, PAWEŁ PRAŁAT‡, AND NICHOLAS WORMALD§

Abstract. A model for *cleaning* a graph with brushes was recently introduced. We consider the minimum number of brushes needed to clean d-regular graphs in this model, focusing on the asymptotic number for random d-regular graphs. We use a degree-greedy algorithm to clean a random d-regular graph on n vertices (with dn even) and analyze it using the differential equations method to find the (asymptotic) number of brushes needed to clean a random d-regular graph using this algorithm (for fixed d). We further show that for any d-regular graph on n vertices at most $n(d+1)/4$ brushes suffice and prove that, for fixed large d, the minimum number of brushes needed to clean a random d-regular graph on n vertices is asymptotically almost surely $\frac{n}{4}(d+o(d))$, thus solving a problem raised in [M.E. Messinger, R.J. Nowakowski, P. Prałat, and N. Wormald, *Cleaning random d-regular graphs with brushes using a degree-greedy algorithm*, in Combinatorial and Algorithmic Aspects of Networking, Lecture Notes in Comput. Sci. 4852, Springer, Berlin-Heidelberg, 2007, pp. 13–26].

Key words. cleaning process, random d-regular graphs, degree-greedy algorithm, differential equations method

AMS subject classification. 05C80

DOI. 10.1137/070703053

1. Introduction.

The cleaning model, introduced in [15, 16], is a combination of the chip-firing game and edge searching on a simple finite graph. (See also [11] where the parallel version of the process is studied.) The brush number of a graph G defined below corresponds to the minimum total imbalance of G which is used in the graph drawing theory. For the starting point of many graph drawing algorithms, a "balanced" ordering of the vertices is required; see, for example, [6] for more.

Initially, every edge and vertex of a graph is *dirty* and a fixed number of brushes start on a set of vertices. At each step, a vertex v and all of its incident edges which are dirty may be *cleaned* if there are at least as many brushes on v as there are incident dirty edges. When a vertex is cleaned, every incident dirty edge is traversed (i.e., cleaned) by one and only one brush, and brushes cannot traverse a clean edge. See Figure 1 for an example of this cleaning process. The initial configuration has only 2 brushes, both at a. The solid edges are dirty, and the dotted edges are clean. The circle indicates which vertex is cleaned next.

The assumption in [16], and taken here, is that *a graph is cleaned when every vertex has been cleaned.* If every vertex has been cleaned, it follows that every edge has been cleaned. It may be that a vertex v has no incident dirty edges at the time

*Received by the editors September 16, 2007; accepted for publication (in revised form) September 2, 2008; published electronically December 19, 2008.
http://www.siam.org/journals/sidma/23-1/70305.html
†Schools of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel (noga@post.tau.ac.il). This research was supported in part by the Israel Science Foundation, by a USA-Israel BSF grant, and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.
‡Department of Mathematics and Statistics, Dalhousie University, Halifax B3H 3P6, NS, Canada (pralat@mathstat.dal.ca). This research was partially supported by grants from NSERC and MITACS.
§Department of Combinatorics and Optimization, University of Waterloo, Waterloo N2L 3G1, ON, Canada (nickwor@math.uwaterloo.ca). This research was supported by the Canada Research Chairs Program and NSERC.

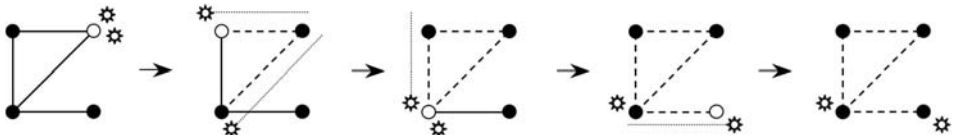


FIG. 1. *An example of the cleaning process.*

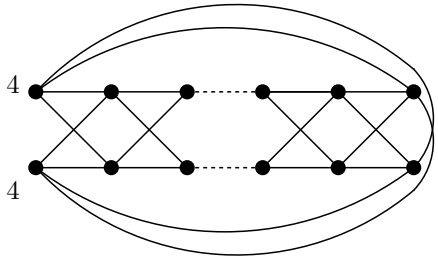


FIG. 2. *An example of the cleaning process for a* 4*-regular graph requiring* 8 *brushes.*

it is cleaned, in which case no brushes move from v. Although this viewpoint might seem unnatural, it simplified much of the analysis in [16].

In this paper, we are interested in the asymptotic number of brushes needed to clean d-regular and mainly random d-regular (finite, simple) graphs. At one extreme, the graph could consist of disjoint copies of K_{d+1}. From [16], K_{d+1} requires essentially $d^2/4$ brushes so that the whole graph requires approximately $nd/4$. At the lower end, if d is even, then a ring of bipartite graphs $K_{d/2,d/2}$ chained together (see Figure 2 for the case $d = 4$) requires only $d^2/2$ brushes regardless of the number of vertices (by placing d brushes at each of $d/2$ vertices on one "level" and working around the ring). If d is odd, then every vertex has at least one brush in either the original or the final configuration (see [16] for more details) so that a graph on n vertices requires at least $n/2$ brushes.

In section 2 we introduce the formal definitions for the cleaning process and also include a description of the pairing model which is used in the results on random regular graphs, instead of working directly in the uniform probability space.

In section 3 we describe some general upper and lower bounds for the minimum number of brushes needed to clean a graph and show, in particular, that for any d-regular graph on n vertices, $n(d+1)/4$ brushes suffice if d is odd and $\frac{n}{4}(d+1-\frac{1}{d+1})$ brushes suffice if d is even. These bounds are tight. We also show that for random d-regular graphs on n vertices the minimum number of brushes needed is, asymptotically almost surely (a.a.s.), at least $\frac{n}{4}(d - O(\sqrt{d}))$.

Section 4 concerns random d-regular graphs. Most of the results in this section form an extended version of the conference paper [17]. We first observe that if $d = 2$, then the brush number of a random d-regular graph on n vertices is a.a.s. $(1 + o(1)) \log n$; for $d = 3$, the brush number is equal to $n/2 + 2$ a.a.s.; for $d = 4$, $(1 + o(1))n/3$ brushes are enough to clean a graph a.a.s. and for $d = 5$, roughly $0.644n$. In order to get an asymptotically almost sure upper bound on the brush number we use a degree-greedy algorithm [22] to clean the graph and then use the differential equation method, studied in [25], to find the asymptotic number of brushes required. We also consider the case of large d and show that the typical brush number in this case is roughly $nd/4$, thus solving a problem raised in [17].

We conclude with a few open problems.

2. Definitions. The following cleaning algorithm and terminology was recently introduced in [16].

Formally, at each step t, $\omega_t(v)$ denotes the number of brushes at vertex v ($\omega_t : V \to \mathbb{N} \cup \{0\}$) and D_t denotes the set of dirty vertices. An edge $uv \in E$ is dirty if and only if both u and v are dirty: $\{u, v\} \subseteq D_t$. Finally, let $D_t(v)$ denote the number of dirty edges incident to v at step t:

$$D_t(v) = \begin{cases} |N(v) \cap D_t| & \text{if } \ v \in D_t, \\ 0 & \text{otherwise.} \end{cases}$$

DEFINITION 2.1. *The* cleaning process $\mathfrak{P}(G, \omega_0) = \{(\omega_t, D_t)\}_{t=0}^{T}$ *of an undirected graph* $G = (V, E)$ *with an* initial configuration of brushes ω_0 *is as follows:*

(0) *Initially, all vertices are dirty:* $D_0 = V$*; set* $t := 0$.
(1) *Let* α_{t+1} *be any vertex in* D_t *such that* $\omega_t(\alpha_{t+1}) \geq D_t(\alpha_{t+1})$. *If no such vertex exists, then stop the process, set* $T = t$, *and return the* cleaning sequence $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_T)$, *the* final set of dirty vertices D_T, *and the* final configuration of brushes ω_T.
(2) *Clean* α_{t+1} *and all dirty incident edges by moving a brush from* α_{t+1} *to each dirty neighbor. More precisely,* $D_{t+1} = D_t \setminus \{\alpha_{t+1}\}$, $\omega_{t+1}(\alpha_{t+1}) = \omega_t(\alpha_{t+1}) - D_t(\alpha_{t+1})$, *and for every* $v \in N(\alpha_{t+1}) \cap D_t$, $\omega_{t+1}(v) = \omega_t(v) + 1$ *(the other values of* ω_{t+1} *remain the same as in* ω_t*).*
(3) $t := t + 1$, *and go back to* (1).

Note that for a graph G and initial configuration ω_0 the cleaning process can return different cleaning sequences and final configurations of brushes; consider, for example, an isolated edge uv and $\omega_0(u) = \omega_0(v) = 1$. It has been shown (see Theorem 2.1 in [16]), however, that the final set of dirty vertices is determined by G and ω_0. Thus, the following definition is natural.

DEFINITION 2.2. *A graph* $G = (V, E)$ can be cleaned *by the initial configuration of brushes* ω_0 *if the cleaning process* $\mathfrak{P}(G, \omega_0)$ *returns an empty final set of dirty vertices* ($D_T = \emptyset$).

Let the brush number $b(G)$ *be the minimum number of brushes needed to clean* G, *that is,*

$$b(G) = \min_{\omega_0 : V \to \mathbb{N} \cup \{0\}} \left\{ \sum_{v \in V} \omega_0(v) : G \text{ can be cleaned by } \omega_0 \right\}.$$

Similarly, $b_\alpha(G)$ *is defined as the minimum number of brushes needed to clean* G *using the cleaning sequence* α.

It is clear that for every cleaning sequence α, $b_\alpha(G) \geq b(G)$ and $b(G) = \min_\alpha b_\alpha(G)$. (The last relation can be used as an alternative definition of $b(G)$.) In general, it is difficult to find $b(G)$, but $b_\alpha(G)$ can be easily computed. For this, it seems better not to choose the function ω_0 in advance, but to run the cleaning process in the order α, and compute the initial number of brushes needed to clean a vertex. We can adjust ω_0 along the way

$$\omega_0(\alpha_{t+1}) = \max\{2D_t(\alpha_{t+1}) - \deg(\alpha_{t+1}), 0\}, \quad \text{for } t = 0, 1, \ldots, |V| - 1, \tag{1}$$

since that is the number of brushes we have to add over and above what we get for free.

Our main results refer to the probability space of random d-regular graphs with uniform probability distribution. This space is denoted by $\mathcal{G}_{n,d}$, and asymptotics (such as a.a.s.) are for $n \to \infty$ with $d \geq 2$ fixed and n even if d is odd.

Instead of working directly in the uniform probability space of random regular graphs on n vertices $\mathcal{G}_{n,d}$, we use the *pairing model* of random regular graphs, first introduced by Bollobás [7], which is described next. Suppose that dn is even, as in the case of random regular graphs, and consider dn points partitioned into n labeled buckets $v_1, v_2, \ldots, v_n$ of d points each. A *pairing* of these points is a perfect matching into $dn/2$ pairs. Given a pairing P, we may construct a multigraph $G(P)$, with loops allowed, as follows: the vertices are the buckets $v_1, v_2, \ldots, v_n$, and a pair $\{x, y\}$ in P corresponds to an edge $v_i v_j$ in $G(P)$ if x and y are contained in the buckets v_i and v_j, respectively. It is an easy fact that the probability of a random pairing corresponding to a given simple graph G is independent of the graph, and hence the restriction of the probability space of random pairings to simple graphs is precisely $\mathcal{G}_{n,d}$. Moreover, it is well known that a random pairing generates a simple graph with probability asymptotic to $e^{(1-d^2)/4}$ depending on d so that any event holding a.a.s. over the probability space of random pairings also holds a.a.s. over the corresponding space $\mathcal{G}_{n,d}$. For this reason, asymptotic results over random pairings suffice for our purposes. One of the advantages of using this model is that the pairs may be chosen sequentially so that the next pair is chosen uniformly at random over the remaining (unchosen) points. For more information on this model, see [23].

3. Bounds.

3.1. Lower bounds.

When a graph G is cleaned using the cleaning process described in Definition 2.1, each edge of G is traversed exactly once and by exactly one brush.

Note that no brush may return to a vertex it has already visited, motivating the following definition.

DEFINITION 3.1. *The* brush path *of a brush b is the path formed by the set of edges cleaned by b.*

By definition, G can be decomposed into $b_\alpha(G)$ brush paths. (Since no brush can stay at its initial vertex in the minimal brush configuration, these paths each have at least one edge.) Thus, the minimum number of paths into which a graph G can be decomposed yields a lower bound for $b(G)$. This is only a lower bound because some path decompositions would not be valid in the cleaning process. For example, K_4 can be decomposed into two edge-disjoint paths, but $b(K_4) = 4$.

In any path decomposition, every vertex of odd degree in a graph G will be the end point of (at least) one path. This leads to a natural lower bound for $b(G)$ since a graph with d_o odd vertices cannot be decomposed into less than $d_o/2$ paths (see [16] for more details).

THEOREM 3.2. *Given initial configuration ω_0, suppose G can be cleaned yielding final configuration ω_T. Then for every vertex v in G with odd degree either $\omega_0(v) > 0$ or $\omega_T(v) > 0$. In particular, $b(G) \geq d_o(G)/2$, where $d_o(G)$ denotes the number of vertices of odd degree.*

The result can be improved a little if there is a lower bound on the vertex degrees (see section 4.3 for details).

Another general lower bound for random d-regular graphs can be obtained as follows. By [16, Theorem 3.2],

$$b(G) \geq \max_j \min_{S \subseteq V, |S|=j} \{jd - 2|E(G[S])|\} = \max_j \min_{S \subseteq V, |S|=j} |E(S, V \setminus S)|, \tag{2}$$

where $E(S, V \setminus S)$ is the set of all edges between S and its complement and $E(G[S])$ is the set of all edges in the induced subgraph of G on S. The proof is a simple corollary of the fact that the minimum above is a lower bound on the number of edges going from the first j vertices cleaned to elsewhere in the graph. So suppose that x and y are functions of n such that the expected number $S(x, y)$ of sets S of xn vertices in $G \in \mathcal{G}_{n,d}$ with yn edges to the complement $V(G) \setminus S$ is $o(1)$. Then this theorem, together with the first moment principle, gives that the brush number is a.a.s. at least yn.

In order to find optimal values of x and y we use the pairing model. (This is essentially the same argument used by Bollobás [9] to obtain a lower bound on the isoperimetric number of random regular graphs, but since it is slightly simpler for our purposes and we obtain a slightly different conclusion, we include the argument.) It is clear that

$$S(x,y) = \binom{n}{xn}\binom{xdn}{yn} M(xdn - yn) \binom{(1-x)dn}{yn} (yn)! M((1-x)dn - yn)/M(dn),$$

where $M(i)$ is the number of perfect matchings on i vertices, that is,

$$M(i) = \frac{i!}{(i/2)!2^{i/2}}.$$

After simplification we get

$$S(x,y) = \frac{n!(xdn)!((1-x)dn)!(dn/2)!2^{yn}}{(xn)!((1-x)n)!(yn)!((xd-y)n/2)!(((1-x)d-y)n/2)!(dn)!}.$$

Using Stirling's formula ($n! \sim \sqrt{2\pi n}(n/e)^n$) and taking the exponential part, we obtain

$$\begin{aligned} S(x,y) &\le e^{o(n)} \frac{x^{x(d-1)n}(1-x)^{(1-x)(d-1)n} d^{dn/2}}{y^{yn}(xd-y)^{(xd-y)n/2}((1-x)d-y)^{((1-x)d-y)n/2}} \\ &= e^{-f(x,y,d)n + o(n)}, \end{aligned} \tag{3}$$

where

$$\begin{aligned} f(x,y,d) = {} & x(d-1)\ln x + (1-x)(d-1)\ln(1-x) + 0.5d\ln d - y\ln y \\ & -0.5(xd-y)\ln(xd-y) - 0.5((1-x)d-y)\ln((1-x)d-y). \end{aligned}$$

Thus, if $f(x, y, d) < 0$, then $S(x, y)$ is exponentially small (n large) and the brush number is at least yn. Not surprisingly, the strongest bound is obtained for $x = 1/2$, in which case $f(x, y, d)$ becomes

$$\begin{aligned} & (d-1)\ln(1/2) + (d/2)\ln d - y\ln y - (d/2-y)\ln(d/2-y) \\ & \qquad = -\frac{d}{4}((1+z)\ln(1+z) + (1-z)\ln(1-z)) + \ln 2, \end{aligned}$$

where $y = (d/4)(1-z)$.

It is straightforward to see that this function is decreasing in z for $z \ge 0$. Let l_d/n denote the value of y for which it first reaches 0. Using the full power of Stirling's formula, it is also not difficult to see that we can replace $e^{o(n)}$ by $O(n^{-1})$ in (3). This gives us the following asymptotically almost sure lower bounds l_d for the brush number of the random d-regular graph: $l_4 = 0.22n$, $l_5 = 0.36n$, and $l_6 = 0.52n$. (In

this paper, whenever we quote numerical values for computed constants such as l_d/n, we use three decimal places rounded down for lower bounds and up for upper bounds.)

In Figure 6, the values of l_d/dn have been presented for all d-values up to 100; we have also listed the first 30 and a few more values for higher d in Table 1 (see section 4.6).

To obtain a result useful for all d, it is straightforward to show (since the Taylor expansion of $(1+z)\ln(1+z) + (1-z)\ln(1-z)$ is $z^2 + z^4/6 + \cdots$) that $l_d/n > (d/4)(1 - 2\sqrt{\ln 2}/\sqrt{d})$. This result has the following implication giving a nontrivial lower bound for $d \geq 3$.

COROLLARY 3.3. *For $G \in \mathcal{G}_{n,d}$, a.a.s.*

$$b(G) \geq \frac{dn}{4}\left(1 - \frac{2\sqrt{\ln 2}}{\sqrt{d}}\right).$$

Alternatively, one can use the expansion properties of random d-regular graphs that follow from their eigenvalues to get a similar lower bound.

The adjacency matrix $A = A(G)$ of a given d-regular graph G with n vertices is an $n \times n$ real and symmetric matrix. Thus, the matrix A has n real eigenvalues which we denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. It is known that certain properties of a d-regular graph are reflected in its spectrum, but, since we focus on expansion properties, we are particularly interested in the following quantity: $\lambda = \lambda(G) = \max(|\lambda_2|, |\lambda_n|)$. In other words, λ is the largest absolute value of an eigenvalue other than $\lambda_1 = d$ (for more details, see the general survey [12] about expanders or Chapter 9 of [5]).

The value of λ for random d-regular graphs has been studied extensively. A major result due to Friedman [10] is the following.

LEMMA 3.4 (see [10]). *For every fixed $\varepsilon > 0$ and for $G \in \mathcal{G}_{n,d}$,*

$$\mathbb{P}(\lambda(G) \leq 2\sqrt{d-1} + \varepsilon) = 1 - o(1).$$

The number of edges $|E(S,T)|$ between sets S and T is expected to be close to the expected number of edges between S and T in a random graph of edge density d/n, namely, $d|S||T|/n$. A small λ (or large spectral gap) implies that this deviation is small. The following useful bound is essentially proved in [2] (see also [5]).

LEMMA 3.5 (expander mixing lemma). *Let G be a d-regular graph with n vertices, and set $\lambda = \lambda(G)$. Then for all $S, T \subseteq V$*

$$\left|\,|E(S,T)| - \frac{d|S||T|}{n}\,\right| \leq \lambda\sqrt{|S||T|}.$$

(Note that $S \cap T$ does not have to be empty; in general, $|E(S,T)|$ is defined to be the number of edges between $S \setminus T$ to T plus twice the number of edges that contain only vertices of $S \cap T$.)

For our purpose here it is better to apply a slightly stronger lower estimate for $|E(S, V \setminus S)|$, namely,

$$|E(S, V \setminus S)| \geq \frac{(d-\lambda)|S||V \setminus S|}{n} \tag{4}$$

for all $S \subseteq V$. This is proved in [4]; see also [5].

From (4) and Lemma 3.4 we get immediately the following corollary. (In order to get the second part, it is enough to use (2) with $j = \lfloor n/2 \rfloor$. The second part is only slightly weaker than Corollary 3.3.)

COROLLARY 3.6. *Let $G \in \mathcal{G}_{n,d}$. For every $\varepsilon > 0$, a.a.s. all $S \subseteq V(G)$ satisfy the following condition:*

$$|E(S, V \setminus S)| \geq \frac{(d - 2\sqrt{d-1} - \varepsilon)|S||V \setminus S|}{n}.$$

In particular, a.a.s.

$$b(G) \geq \frac{dn}{4}\left(1 - \frac{2}{\sqrt{d}}\right).$$

Remark. The minimum number of edges in a cut that splits the vertex set of a graph into two equal parts is called its bisection width. In the above arguments we have used it as a lower bound for the brush number of the graph. It is worth noting that the $\frac{2}{\sqrt{d}}$ error term in the lower bound for the bisection width of a d-regular graph on n vertices is tight, up to a constant factor. Indeed, it is shown in [1] that for $n \gg d$ the bisection width of any d-regular graph on n vertices is at most $\frac{nd}{4}(1 - \Omega(\frac{1}{\sqrt{d}}))$.

3.2. A general upper bound. The following result provides an upper bound for the brush number of a general graph.

THEOREM 3.7.

$$b(G) \leq \frac{|E|}{2} + \frac{|V|}{4} - \frac{1}{4} \sum_{v \in V(G), \deg(v) \text{ is even}} \frac{1}{\deg(v) + 1}$$

for any graph $G = (V, E)$.

Proof. Let π be a random permutation of the vertices of G taken with uniform distribution. We clean G according to this permutation to get the value of $b_\pi(G)$ (note that $b_\pi(G)$ is a random variable now). For a vertex $v \in V$, it follows from (1) that we have to assign to v exactly $X(v) = \max\{0, 2N^+(v) - \deg(v)\}$ brushes in the initial configuration, where $N^+(v)$ is the number of neighbors of v that follow it in the permutation (that is, the number of dirty neighbors of v at the time when v is cleaned). The random variable $N^+(v)$ attains each of the values $0, 1, \ldots, \deg(v)$ with probability $1/(\deg(v) + 1)$. Indeed, this follows from the fact that the random permutation π induces a uniform, random permutation on the set of $\deg(v)+1$ vertices consisting of v and its neighbors. Therefore, the expected value of $X(v)$ for even $\deg(v)$, is

$$\frac{\deg(v) + (\deg(v) - 2) + \cdots + 2}{\deg(v) + 1} = \frac{\deg(v) + 1}{4} - \frac{1}{4(\deg(v) + 1)}$$

and for odd $\deg(v)$ it is

$$\frac{\deg(v) + (\deg(v) - 2) + \cdots + 1}{\deg(v) + 1} = \frac{\deg(v) + 1}{4}.$$

Thus, by linearity of expectation,

$$\mathbb{E}b_\pi(G) = \mathbb{E}\left(\sum_{v \in V} X(v)\right) = \sum_{v \in V} \mathbb{E}X(v) = \frac{|E|}{2} + \frac{|V|}{4} - \frac{1}{4} \sum_{v \in V(G), \deg(v) \text{ is even}} \frac{1}{\deg(v) + 1},$$

which means that there is a permutation π_0 such that $b(G) \leq b_{\pi_0}(G) \leq \mathbb{E}b_\pi(G)$, and the assertion holds. ∎

Note that the bound is tight when G is a union of cliques. From this we get immediately the following corollary.

COROLLARY 3.8. *Let $G = (V, E)$ be a d-regular graph on n vertices. If d is even, then*

$$b(G) \le \frac{n}{4}\left(d + 1 - \frac{1}{d+1}\right),$$

and if d is odd, then

$$b(G) \le \frac{n}{4}(d+1).$$

Both bounds are tight for every n and d satisfying $(d+1)|n$, as shown by a disjoint union of complete graphs K_{d+1}.

4. Cleaning random *d*-regular graphs. The differential equations method (described in [25]) is used here to find an upper bound on the number of brushes needed to clean a graph using a degree-greedy algorithm. We consider $d = 2$ first, then state some general results, and apply them to the special cases of $3 \le d \le 5$ before discussing higher values of d.

4.1. 2-regular graphs. Let $Y = Y_n$ be the total number of cycles in a random 2-regular graph on n vertices. Since exactly two brushes are needed to clean one cycle, we need $2Y_n$ brushes in order to clean a 2-regular graph.

We know that the random 2-regular graph is a.a.s. disconnected; by simple calculations one can show that the probability of having a Hamiltonian cycle is asymptotic to $\frac{1}{2}e^{3/4}\sqrt{\pi}n^{-1/2}$ (see, for example, [23]).

We also know that the total number of cycles Y_n is sharply concentrated near $(1/2)\log n$. It is not difficult to see this by generating the random graph sequentially using the pairing model. The probability of forming a cycle in step i is exactly $1/(2n-2i+1)$, so the expected number of cycles is $(1/2)\log n + O(1)$. The variance can be calculated in a similar way. So we get that a.a.s. the brush number for a random 2-regular graph is $(1 + o(1))\log n$.

4.2. *d*-regular graphs ($d \ge 3$)—the general setting. In this subsection, we assume $d \ge 3$ is fixed with dn even. In order to get an asymptotically almost sure upper bound on the brush number, we study an algorithm that cleans random vertices of minimum degree. This algorithm is called *degree-greedy* because the vertex being cleaned is chosen from those with the lowest degree.

We start with a random d-regular graph $G = (V, E)$ on n vertices. Initially, all vertices are dirty: $D_0 = V$. In every step t of the cleaning process, we clean a random vertex α_t, chosen uniformly at random from those vertices with the lowest degree in the induced subgraph $G[D_{t-1}]$, where $D_t = D_{t-1}\setminus\{\alpha_t\}$. In the first step, d brushes are needed to clean a random vertex α_1 (we say that this is "phase zero"). The induced subgraph $G[D_1]$ now has d vertices of degree $d-1$ and $n-d-1$ vertices of degree d. Note that α_1 is a.a.s. the only vertex whose degree in $G[D_t]$ is d at the time of cleaning. Indeed, if α_t ($t \ge 2$) has degree d in $G[D_{t-1}]$, then $G[D_{t-1}]$ consists of some connected components of G, and thus G is disconnected. It was proven independently in [8, 24] that, for constant d, G is disconnected with probability $o(1)$ (this also holds when d is growing with n, as shown in [14]).

In the second step, $d-2$ brushes are needed to clean a random vertex α_2 of degree $d-1$. Typically, in the third step, a vertex of degree $d-1$ is cleaned, and in each

subsequent step, a vertex of degree $d-1$ in $G[D_t]$ is cleaned until some vertex of degree $d-2$ is produced in the subgraph induced by the set of dirty vertices. After cleaning the first vertex of degree $d-2$, we typically return to cleaning vertices of degree $d-1$, but after some more steps of this type we may clean another vertex of degree $d-2$. When vertices of degree $d-1$ become plentiful, vertices of lower degree are more commonly created, and these hiccups occur more often. When vertices of degree $d-2$ take over the role of vertices of degree $d-1$, we say (informally!) that the first phase ends, and we begin the second phase. In general, in the kth phase a mixture of vertices of degree $d-k$ and $d-k-1$ are cleaned.

During the kth phase there are, in theory, two possible endings. It can happen that the vertices of degree $d-k$ are becoming so common that the vertices of degree $d-k-1$ start to explode (in which case we move to the next phase). It is also possible that the ones of degree $d-k+1$ are getting so rare that those of degree $d-k$ disappear (in which case the process goes "backwards"). With various initial conditions, either one could occur. However, the numerical solutions of the DEs for $d = 4, 5, \ldots, 100$ support the hypothesis that the degree-greedy process we study never goes "back." In such cases, the remaining vertices are cleaned "for free" (that is, after the crucial phases, only $o(n)$ new brushes are required to finish the process). The details of the following differential equations method have been omitted but can be found in [22].

For $0 \le i \le d$, let $Y_i = Y_i(t)$ denote the number of vertices of degree i in $G[D_t]$. (Note that $Y_0(t) = n - t - \sum_{i=1}^{d} Y_i(t)$ so $Y_0(t)$ does not need to be calculated, but it is useful in the discussion.) Let $S(t) = \sum_{l=1}^{d} lY_l(t)$, and for any statement A, let δ_A denote the Kronecker delta function

$$\delta_A = \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$$

It is not difficult to see that

$$\begin{aligned} &\mathbb{E}(Y_i(t) - Y_i(t-1) \mid G[D_{t-1}] \ \wedge \ \deg_{G[D_{t-1}]}(\alpha_t) = r) \\ &\qquad = f_{i,r}((t-1)/n, Y_1(t-1)/n, Y_2(t-1)/n, \ldots, Y_d(t-1)/n) \\ &\qquad = -\delta_{i=r} - r\frac{iY_i(t-1)}{S(t-1)} + r\frac{(i+1)Y_{i+1}(t-1)}{S(t-1)}\delta_{i+1\le d} \end{aligned} \tag{5}$$

for $i, r \in [d]$ such that $Y_r(t) > 0$. Indeed, α_t has degree r and hence the term $-\delta_{i=r}$. When a pair of points in the pairing model is exposed, the probability that the other point is in a bucket of degree i (that is, the bucket contains i unchosen points) is asymptotic to $iY_i(t-1)/S(t-1)$. Thus, $riY_i(t-1)/S(t-1)$ stands for the expected number of the r buckets found adjacent to α_t which have degree i. This contributes negatively to the expected change in Y_i, while buckets of degree $i+1$ which are reached contribute positively (of course, only if this type of vertices (buckets) exists in a graph; thus $\delta_{i+1\le d}$). This explains (5).

Suppose that, at some step t of the phase k, cleaning a vertex of degree $d-k$ creates, in expectation, β_k vertices of degree $d-k-1$ and cleaning a vertex of degree $d-k-1$ decreases, in expectation, the number of vertices of degree $d-k-1$ by τ_k. After cleaning a vertex of degree $d-k$, we expect to then clean (on average) β_k/τ_k vertices of degree $d-k-1$. Thus, in phase k, the proportion of steps which clean vertices of degree $d-k$ is $1/(1+\beta_k/\tau_k) = \tau_k/(\beta_k+\tau_k)$. If τ_k falls below zero, vertices of degree $d-k-1$ begin to build up and do not decrease under repeated cleaning vertices of this type, and we move to the next phase.

From (5) it follows that

$$\beta_k = \beta_k(x, y_1, y_2, \dots, y_d) = f_{d-k-1,d-k}(x, y_1, y_2, \dots, y_d) = f_{d-k-1,d-k}(x, \mathbf{y}),$$
$$\tau_k = \tau_k(x, y_1, y_2, \dots, y_d) = -f_{d-k-1,d-k-1}(x, y_1, y_2, \dots, y_d) = -f_{d-k-1,d-k-1}(x, \mathbf{y}),$$

where $x = t/n$ and $y_i(x) = Y_i(t)/n$ for $i \in [d]$. This suggests (see [25] for more information on the differential equations method) the following system of differential equations:

$$\frac{dy_i}{dx} = F(x, \mathbf{y}, i, k),$$

where

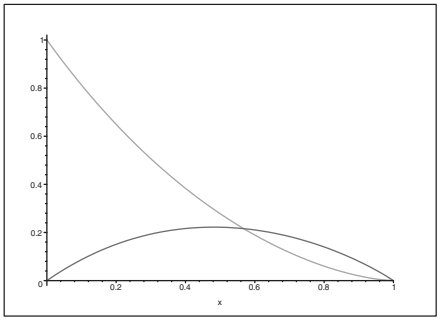
$$F(x, \mathbf{y}, i, k) = \begin{cases} \dfrac{\tau_k}{\beta_k + \tau_k} f_{i,d-k}(x, \mathbf{y}) + \dfrac{\beta_k}{\beta_k + \tau_k} f_{i,d-k-1}(x, \mathbf{y}) & \text{for } k \le d-2, \\ f_{i,1}(x, \mathbf{y}) & \text{for } k = d-1. \end{cases}$$

At this point we may formally define the interval $[x_{k-1}, x_k]$ to be phase k, where the termination point x_k is defined as the infimum of those $x > x_k$ for which at least one of the following holds: $\tau_k \le 0$ and $k < d-1$; $\tau_k + \beta_k = 0$ and $k < d-1$; $y_{d-k} \le 0$. Using final values $y_i(x_k)$ in phase k as initial values for phase $k+1$, we can repeat the argument inductively moving from phase to phase starting from phase 1 with obvious initial conditions $y_d(0) = 1$ and $y_i(0) = 0$ for $0 \le i \le d-1$.

The general result [22, Theorem 1] studies a deprioritized version of degree-greedy algorithms, which means that the vertices are chosen to process in a slightly different way, not always the minimum degree, but usually a random mixture of two degrees. Once a vertex is chosen, it is treated the same as in the degree-greedy algorithm. The variables Y are defined in an analogous manner. The hypotheses of the theorem are mainly straightforward to verify but require several inequalities involving derivatives to hold at the termination of phases for the full rigorous conclusion to be obtained. However, in practice, the equations are simply solved numerically in order to find the points x_k, since a fully rigorous bound is not obtained unless one obtains strict inequalities on the values of the solutions. The conclusion is that, for a certain algorithm using a deprioritized "mixture" of the steps of the degree-greedy algorithm, with variables Y_i defined as above, we have that a.a.s.

$$Y_i(t) = n y_i(t/n) + o(n)$$

for $1 \le i \le d$ for phases $k = 1, 2, \dots, m$, where m denotes the smallest k for which either $k = d-1$ or any of the termination conditions for phase k hold at x_k apart from $x_k = \inf\{x > x_{k-1} : \tau_k \le 0\}$. We omit all details, pointing the reader to [22] and the general survey [25] about the differential equations method, which is a main tool in proving [22, Theorem 1]. In addition, the theorem gives information on an auxiliary variable such as, of importance to our present application, the number of brushes used. Instead of quoting this precisely, we use it merely as justification for being able to use the above equations as if they applied to the greedy algorithm. (This is no doubt the case, but it is not actually proved in [22]. Instead, we know that they apply in the limit to a sequence of algorithms that use the steps of the degree-greedy algorithm.) The solution to the relevant differential equations for $d = 3$ is shown in Figure 3.



(a) 3-regular graph, phase 1

FIG. 3. *Solution to the differential equations.*

In the kth phase a mixture of vertices of degree $d-k$ and $d-k-1$ is cleaned. Since $\max\{2l-d,0\}$ brushes are needed to clean a vertex of degree l (see (1)), we need

$$u_d^k = (1+o(1))n\left(\max\{d-2k,0\}\int_{x_{k-1}}^{x_k}\frac{\tau_k}{\tau_k+\beta_k}dx + \max\{d-2k-2,0\}\int_{x_{k-1}}^{x_k}\frac{\beta_k}{\tau_k+\beta_k}dx\right)$$

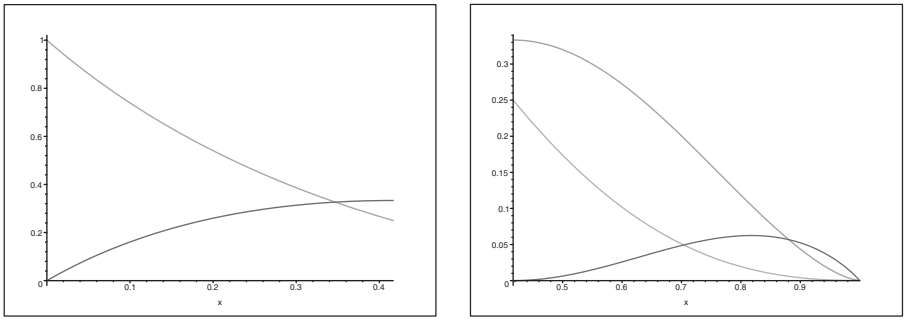
brushes in phase k. Thus, the total number of brushes needed to clean a graph using the degree-greedy algorithm is a.a.s. equal to

$$\begin{aligned} u_d &= \sum_{k=1}^{\lfloor (d-1)/2\rfloor} u_d^k + o(n) \\ &= (1+o(1))n\left(\sum_{k=1}^{\lfloor (d-1)/2\rfloor}\left((d-2k-2)(x_k-x_{k-1})+2\int_{x_{k-1}}^{x_k}\frac{\tau_k}{\tau_k+\beta_k}dx\right) + \delta_{d \text{ is odd}}\int_{x_{k-1}}^{x_k}\frac{\beta_k}{\tau_k+\beta_k}dx\right). \end{aligned}$$

(We assume here that the solutions of the DEs proceed in the way we presume, that is, with no "reversion" to earlier phases. This implies that only $o(n)$ new brushes are required for the remaining phases.)

4.3. 3-regular graphs. Let $G=(V,E)$ be any 3-regular graph on n vertices. The first vertex cleaned must start three brush paths, the last one terminates three brush paths, and all other vertices must start or finish at least one brush path, so the number of brush paths is at least $n/2+2$.

The result mentioned above can be shown to result in an upper bound of $n/2+o(n)$ for the brush number of a random 3-regular (i.e., cubic) graph. We do not provide details because of the following stronger result. It is known [21] that a random 3-regular graph a.a.s. has a Hamilton cycle. The edges not in a Hamilton cycle must form a perfect matching. Such a graph can be cleaned by starting with three brushes at one vertex and moving along the Hamilton cycle with one brush, introducing one



(a) 4-regular graph, phase 1

(b) 4-regular graph, phase 2 (vertices are cleaned "for free")

FIG. 4. *Solution to the differential equations.*

new brush for each edge of the perfect matching. Hence the brush number of a random 3-regular graph with n vertices is a.a.s. $n/2 + 2$. Note that this is also the brush number of any cubic Hamiltonian graph on n vertices.

4.4. 4-regular graphs. For 4-regular graphs, to estimate the brush number one has to carefully analyze phase 1 only: we need two brushes to clean vertices of degree 3, but vertices of degree 2 are cleaned "for free." Note that $y_1(x) = y_2(x) = 0$ throughout phase 1. We have the following system of differential equations:

$$\frac{dy_4}{dx} = \frac{-6y_4(x)}{3y_3(x) + 2y_4(x)},$$

$$\frac{dy_3}{dx} = \frac{-3y_3(x) + 4y_4(x)}{3y_3(x) + 2y_4(x)},$$

with the initial conditions $y_4(0) = 1$ and $y_3(0) = 0$. The solution (see Figure 4(a)) to these differential equations is

$$y_4(x) = 5 - 4\sqrt{1+3x} + 3x,$$

$$y_3(x) = \frac{4(-3 + 3\sqrt{1+3x} - 5x + x\sqrt{1+3x})}{2 - \sqrt{1+3x}},$$

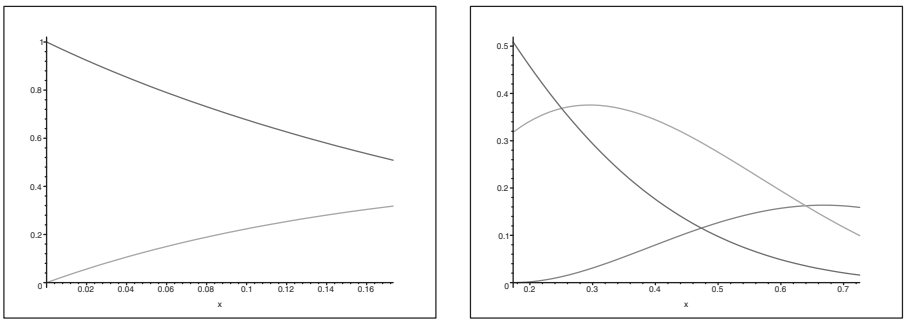
so $\beta_1 = -3 + 3\sqrt{1+3x}$ and $\tau_1 = 3 - 2\sqrt{1+3x}$. Thus, phase 1 finishes at time $t_1 = 5n/12$ ($x_1 = 5/12$ is a root of the equation $\tau_1(x) = 0$), and the number of vertices of degree 3 cleaned during this phase is asymptotic to

$$n \int_0^{5/12} \frac{\tau_1}{\tau_1 + \beta_1} dx = n/6.$$

Since we need 2 brushes to clean one such vertex we get an asymptotically almost sure upper bound of $u_4 = (1 + o(1))n/3$.

The remaining phases can be studied in a similar way, assuring us that no extra brushes are needed. The solution to the relevant differential equations are shown in Figure 4.

On the other hand, it is true that a.a.s. a random 4-regular graph can be decomposed into two edge-disjoint Hamilton cycles [13] and hence four paths.



(a) 5-regular graph, phase 1 (b) 5-regular graph, phase 2

FIG. 5. *Solution to the differential equations.*

Note that the following two problems can be posed in general for any $d \geq 3$.

OPEN PROBLEM 4.1. *Is it true that for the random case it is best to clean lowest degree vertices?*

In other words, if one is going to choose a random vertex of a given degree, is it true that one might as well choose a random vertex of minimum degree?

If Problem 4.1 is proven to be true, then the following problem should be considered. To get the brush number one might (in fact, probably should) choose nonrandom vertices during the cleaning process. But it might be true that a.a.s. one cannot save more than $o(n)$ brushes compared to the greedy algorithm under consideration.

OPEN PROBLEM 4.2. *Is it true that a.a.s. the brush number for a random d-regular graph is* $u_d(1 - o(1))$*?*

4.5. 5-regular graphs. In order to study the brush number for 5-regular graphs yielded by the degree-greedy algorithm, we cannot consider phase 1 only as before; we need 3 brushes to clean vertices of degree 4 but also 1 brush to clean vertices of degree 3. Thus, two phases must be considered.

In phase 1, $y_1(x) = y_2(x) = y_3(x) = 0$, and we have the following system of differential equations:

$$\frac{dy_5}{dx} = \frac{-20y_5(x)}{8y_4(x) + 5y_5(x)},$$

$$\frac{dy_4}{dx} = \frac{-8y_4(x) + 15y_5(x)}{8y_4(x) + 5y_5(x)},$$

with the initial conditions $y_5(0) = 1$ and $y_4(0) = 0$. The numerical solution (see Figure 5(a)) suggests that the phase finishes at time $t_1 = 0.1733n$. The number of brushes needed in this phase is asymptotic to

$$u_5^1 = (1 + o(1)) \left(3n \int_0^{t_1/n} \frac{\tau_1}{\tau_1 + \beta_1} dx + n \int_0^{t_1/n} \frac{\beta_1}{\tau_1 + \beta_1} dx \right)$$

$$= (1 + o(1)) \left(t_1 + 2n \int_0^{t_1/n} \frac{\tau_1}{\tau_1 + \beta_1} dx \right) \approx 0.3180n.$$

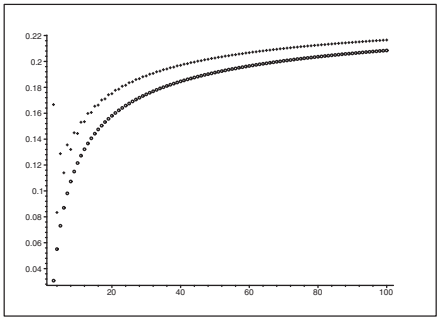


FIG. 6. *A graph of u_d/dn and l_d/dn versus d (from 3 to 100).*

In phase 2, $z_1(x) = z_2(x) = 0$, and we have another system of differential equations

$$\frac{dz_5}{dx} = \frac{-15z_5(x)}{6z_3(x) + 4z_4(x) + 5z_5(x)},$$
$$\frac{dz_4}{dx} = \frac{-3(4z_4 - 5z_5(x))}{6z_3(x) + 4z_4(x) + 5z_5(x)},$$
$$\frac{dz_3}{dx} = \frac{-6z_3(x) + 8z_4(x) - 5z_5(x)}{6z_3(x) + 4z_4(x) + 5z_5(x)},$$

with the initial conditions $z_5(t_1/n) = y_5(t_1/n) = 0.5088$, $z_4(t_1/n) = y_4(t_1/n) = 0.3180$, and $z_3(t_1/n) = 0$. The numerical solution (see Figure 5(b)) suggests that the phase finishes (approximately) at time $t_2 = 0.7257n$. The number of brushes needed in this phase is asymptotic to (the numerical solution)

$$u_5^2 = (1 + o(1))n \int_{t_1/n}^{t_2/n} \frac{\tau_2}{\tau_2 + \beta_2} dx \approx 0.3259n.$$

(Note that there is no $\beta_2/(\tau_2 + \beta_2)$ term this time; each vertex of degree 2 receives 3 extra brushes from already cleaned neighbors and thus can be cleaned "for free.") Finally, we get an asymptotically almost sure upper bound of $u_5 = u_5^1 + u_5^2 \approx 0.6439n$.

4.6. *d*-regular graphs of higher order. Note that the lower bound for $d = 4$ (see section 3.1) will be considerably lower than the lower bound of $n/2 + 2$ for $d = 3$, whereas the upper bound we have been discussing is the same degree-greedy algorithm in all cases. However, the upper bound is also sensitive to the parity of d. For the 4-regular case, vertices of degree 2 are processed "for free," and so one really worries only about degree 3 vertices, and there are fewer of those processed than degree 2 vertices when $d = 3$. But it seems that the parity of d does not greatly affect the value of u_d/n for d big enough (see Figure 6 and Table 1).

In Figure 6, the values of l_d/dn (see section 3.1 for more details about the lower bound) and u_d/dn have been presented for all d-values up to 100, although we have only listed the first 30 and a few more values for higher d in Table 1. (To save effort, the values for $d > 100$ are based on the hypothesis mentioned near the start of section 4.2 that there is no contribution from phases after the $\lfloor (d-1)/2 \rfloor$th one.) The computations presented in the paper were performed by using Maple$^{\text{TM}}$ [18]. The worksheets can be found at the following address: http://www.mathstat.dal.ca/~pralat/.

TABLE 1
Approximate upper and lower bounds on the brush number.

d	l_d/n	u_d/n	d	l_d/n	u_d/n	d	l_d/n	u_d/n	d	l_d/n	u_d/n
3	0.0922	0.500	13	1.77	2.08	23	3.77	4.16	99	20.6	21.5
4	0.220	0.334	14	1.96	2.25	24	3.98	4.36	100	20.8	21.7
5	0.365	0.644	15	2.16	2.49	25	4.18	4.59	149	32.1	33.2
6	0.521	0.684	16	2.35	2.67	26	4.39	4.80	150	32.4	33.5
7	0.686	0.949	17	2.55	2.90	27	4.60	5.03	199	43.8	45.1
8	0.858	1.06	18	2.75	3.08	28	4.81	5.23	200	44.1	45.3
9	1.03	1.31	19	2.95	3.32	29	5.02	5.46	249	55.6	57.0
10	1.21	1.45	20	3.16	3.51	30	5.23	5.67	250	55.9	57.3
11	1.39	1.69	21	3.36	3.74	31	5.44	5.90	299	67.5	69.0
12	1.58	1.85	22	3.56	3.93	32	5.66	6.11	300	67.7	69.3

In [17] the following open question was asked, "does $\lim_{d\to\infty} u_d/dn$ exist?" (Open Problem 3), and it was conjectured that there is a constant c such that the brush number is asymptotically cdn (Open Problem 4). The next theorem settles both questions.

THEOREM 4.3. *The brush number of a random d-regular graph is a.a.s. $\frac{n}{4}(d + o(d))$. Moreover, $\lim_{d\to\infty} u_d/dn = 1/4$, that is, for large d, the degree-greedy algorithm a.a.s. achieves the optimal number of brushes up to a lower order term.*

Proof. The first part of the theorem follows from Corollary 3.3 (or Corollary 3.6) and Corollary 3.8, which show that if $G \in \mathcal{G}_{n,d}$, then a.a.s.

$$\frac{dn}{4}\left(1 - \frac{2\sqrt{\ln 2}}{\sqrt{d}}\right) \le b(G) \le \frac{n(d+1)}{4}.$$

The upper bound here can in fact be slightly improved, as shown in Theorem 4.4 below.

It remains to estimate the performance of the degree-greedy algorithm. Let $d > 2$ be an integer, and let $G \in \mathcal{G}_{n,d}$, as before. It follows from Lemmas 3.4 and 3.5 that a.a.s. for all $m \in \{0, 1, \ldots, n-1\}$ and all sets $X \subseteq V$ with $|X| = m$,

$$|E(G[V \setminus X])| \le \frac{(n-m)^2 d}{2n} + \frac{1}{2} 2\sqrt{d}(n-m)$$

since the number of edges inside $G[V \setminus X]$ is $|E(V \setminus X, V \setminus X)|/2$. So the average degree of $G[V \setminus X]$ (and thus the minimum degree as well) is at most

$$\xi_m = \min\left\{\frac{(n-m)d}{n} + 2\sqrt{d}, d\right\}.$$

Thus, using (1) we get that a.a.s. the number of brushes used by the degree-greedy algorithm is at most

$$\sum_{m=0}^{n-1} \max\{2\xi_m - d, 0\} \le \frac{dn}{4} + O\left(\sqrt{d}n\right).$$

It follows, by Corollary 3.3, that for large d the greedy algorithm achieves, a.a.s., essentially the optimum number of brushes. This completes the proof of the theorem. ∎

The numerical values of the upper bound following from the degree-greedy algorithm suggest that the brush number of a random d-regular graph is a.a.s. smaller than

$dn/4$ for every $d \geq 3$. This fact can be proved by combining the basic idea in the proof of Theorem 3.7 with some known properties of random d-regular graphs. Indeed, the bound in Theorem 3.7 holds for every d-regular graph, and for a random d-regular graph G one can slightly improve the result as follows. It is known (see [23]) that, for the purpose of proving statements a.a.s., such a random graph can be viewed as the multigraph formed from the union of a Hamilton cycle and a random $(d-2)$-regular graph G' on the same vertex set. (The probability of multiple edges being created is bounded away from 1, and the resulting graph, conditional upon no multiple edges, is contiguous to a random d-regular graph. Indeed, Molloy and Reed [19] exploited this fact in a way related to our argument here.) This is equivalent to taking a fixed Hamilton cycle, together with a random $(d-2)$-regular graph G', and permuting its vertices randomly by a permutation π. Therefore, if we clean this multigraph according to the order of the Hamilton cycle, which we denote by $1, 2, \ldots, n$, the edges of G' will be cleaned according to a random permutation. We can thus apply the estimate proved in Corollary 3.8 and conclude that the *expected* number of brushes needed is at most the bound given in that corollary for $(d-2)$-regular graph, plus 2 additional brushes needed to be placed in the first vertex in order to start the process; one of them will keep going along the Hamilton cycle, cleaning all of its edges, and the other one will clean the edge $1, n$ and stay in vertex n until the end of the process. This implies that when G is generated by taking a Hamilton cycle, and a random $(d-2)$-regular G' permuted randomly on the cycle, the expected number of brushes when cleaning along the cycle is at most $2 + \frac{n}{4}(d - 1 - \frac{1}{d-1})$ when d is even and at most $2 + \frac{n}{4}(d-1)$ when d is odd.

However, this is only a bound for the expectation, whereas we need to get an estimate that holds a.a.s. This can be done using a standard martingale argument together with the fact that if we change the permutation π by a single transposition, the number of brushes needed when cleaning along the Hamilton cycle changes by at most $O(d)$ (see, e.g., [3] for a similar argument). Alternatively, since in the random pairing corresponding to G' the number of brushes changes by at most $O(1)$ if two pairs are "switched," [23, Theorem 2.19] immediately implies that a.a.s. the number of brushes required does not deviate from the expectation by more than $O(w(n)\sqrt{n})$, where $w(n)$ is any function tending to infinity with n. We have thus proved the following.

THEOREM 4.4. *Let G be a random d-regular graph on n vertices, where $d \geq 3$. Then, a.a.s., if d is even,*

$$b(G) \leq \frac{n}{4}\left(d - 1 - \frac{1}{d-1}\right)(1 + o(1))$$

and if d is odd, then

$$b(G) \leq \frac{n}{4}(d-1)(1 + o(1)).$$

Note that the numerical bounds obtained using the degree-greedy algorithm appearing in Table 1 are a little stronger than the general one obtained here.

We also note that the estimates in Theorem 4.4 can be further improved by introducing greedy steps into the proof. Instead of simply cleaning along the cycle, one may swap the order of cleaning vertices if such a swap will save a brush, for example, if a vertex has more dirty edges than the next one around the cycle. We do not elaborate on this since, although simple arguments like this will give improvements

that can be described for general d, it seems likely that, carrying the argument as far as possible, one would arrive at the degree-greedy algorithm in any case.

4.7. Variants. We conclude with a few additional open problems.

OPEN PROBLEM 4.5. *What is the brush number for the binomial random graphs $G(n,p)$? What is a lower/upper bound? How about other random graph models, for example, models that give power law degree distribution or d-regular graphs generated by the d-process?*

It is not difficult to show that $b(G) = (1+o(1))pn^2/4$ for $G \in G(n,p)$ and $p > \omega(n)/n$, where $\omega(n)$ is any function tending to infinity. Indeed, in order to get an upper bound it is enough to use Theorem 3.7 since the number of edges is well concentrated around $p\binom{n}{2}$. To get a lower bound, one can show that the expected number of sets of size $\lfloor n/2 \rfloor$ with less than $(1-1/\omega^{1/3}(n))pn^2/4$ edges to its complement is tending to zero as n tends to infinity. The problem of determining the behavior of the brush number for sparser random graphs seems more difficult and has been discussed in [20].

Another version of the cleaning process was introduced in [15]. In this version, when a vertex is cleaned multiple brushes are allowed to traverse each dirty edge. Thus, the brush number $B(G)$ of this generalized version is at most the original one $b(G)$. As before, one can study the behavior of the degree-greedy algorithm to get an asymptotically almost sure upper bound on the generalized brush number. It is clear that there is no point to introduce more brushes in the initial configuration than is required to continue the process (they can be always introduced later when there is need for that). Therefore, the same number of brushes is required for the first $\lfloor (d-1)/2 \rfloor - 1$ phases in both original and generalized versions of the process. During the phase $\lfloor (d-1)/2 \rfloor$ for d even, 2 extra brushes are needed to clean vertex of degree $d/2+1$ in $G[D_t]$, but vertices of degree $d/2$ are cleaned "for free." Since no brush "gets stuck" during this phase in the original model (exactly one brush traverses each edge in the generalized one) and vertices in the last phases are cleaned "for free" (in both models), the upper bounds of the brush numbers are exactly the same. The situation is different when d is odd. During the phase $(d-1)/2$ we can (and should) move two brushes when a vertex of degree $(d-1)/2$ is cleaned and try to save some brushes, but the following is still open.

OPEN PROBLEM 4.6.

- *Is it true that for $G \in \mathcal{G}_{n,d}$, d even, $b(G) - B(G) = o(n)$ a.a.s.?*
- *Is it true that for $G \in \mathcal{G}_{n,d}$, d odd, $b(G) - B(G) = \Theta(n)$ a.a.s.? How far apart are they?*

REFERENCES

[1] N. ALON, *On the edge-expansion of graphs*, Combin. Probab. Comput., 6 (1997), pp. 145–152.

[2] N. ALON AND F.R.K. CHUNG, *Explicit construction of linear sized tolerant networks*, Discrete Math., 72 (1988), pp. 15–19.

[3] N. ALON AND G. GUTIN, *Properly colored Hamilton cycles in edge colored complete graphs*, Random Structures Algorithms, 11 (1997), pp. 179–186.

[4] N. ALON AND V.D. MILMAN, λ_1*, isoperimetric inequalities for graphs and superconcentrators*, J. Combin. Theory Ser. B, 38 (1985), pp. 73–88.

[5] N. ALON AND J.H. SPENCER, *The Probabilistic Method*, Wiley, New York, 1992 (2nd ed., 2000).

[6] T. BIEDL, T. CHAN, Y. GANJALI, M. HAJIAGHAYO, AND D. WOOD, *Balanced vertex–orderings of graphs*, Discrete Appl. Math., 148 (2005), pp. 27–48.

[7] B. BOLLOBÁS, *A probabilistic proof of an asymptotic formula for the number of labelled regular graphs*, European J. Combin., 1 (1980), pp. 311–316.

[8] B. Bollobás, *Random graphs*, in Combinatorics, London Math. Soc. Lecture Note Ser. 52, H.N.V. Temperley, ed., Cambridge University Press, Cambridge, 1981, pp. 80–102.
[9] B. Bollobás, *The isoperimetric number of random regular graphs*, European J. Combin., 9 (1984), pp. 241–244.
[10] J. Friedman, *A proof of Alon's second eigenvalue conjecture*, Mem. Amer. Math. Soc., to appear.
[11] S. Gaspers, M.E. Messinger, R. Nowakowski, and P. Pralat, *Parallel cleaning of a network with brushes*, Discrete Appl. Math., submitted.
[12] S. Hoory, N. Linial, and A. Wigderson, *Expander graphs and their applications*, Bull. Amer. Math. Soc. (N.S.), 43 (2006), pp. 439–561.
[13] J.H. Kim and N.C. Wormald, *Random matchings which induce Hamilton cycles and Hamiltonian decompositions of random regular graphs*, J. Combin. Theory Ser. B, 81 (2001), pp. 20–44.
[14] T. Łuczak, *Sparse random graphs with a given degree sequence*, in Random Graphs, Vol. 2, A. Frieze and T. Łuczak, eds., Wiley, New York, 1992, pp. 165–182.
[15] S. McKeil, *Chip Firing Cleaning Processes*, MSc thesis, Dalhousie University, Halifax, NS, Canada, 2007.
[16] M.E. Messinger, R.J. Nowakowski, and P. Prałat, *Cleaning a network with brushes*, Theoret. Comput. Sci., 399 (2008), pp. 191–205.
[17] M.E. Messinger, R.J. Nowakowski, P. Prałat, and N. Wormald, *Cleaning random d-regular graphs with brushes using a degree-greedy algorithm*, in Combinatorial and Algorithmic Aspects of Networking, Lecture Notes in Comput. Sci. 4852, Springer, Berlin, Heidelberg, 2007, pp. 13–26.
[18] M.B. Monagan, K.O. Geddes, K.M. Heal, G. Labahn, S.M. Vorkoetter, J. McCarron, and P. DeMarco, *Maple* 10 *Programming Guide*, Maplesoft, Waterloo ON, Canada, 2005.
[19] M. Molloy and B. Reed, *The dominating number of a random cubic graph*, Random Structures Algorithms, 7 (1995), pp. 209–221.
[20] P. Prałat, *Cleaning random graphs with brushes*, Australas. J. Combin., to appear.
[21] R.W. Robinson and N.C. Wormald, *Almost all cubic graphs are Hamiltonian*, Random Structures Algorithms, 3 (1992), pp. 117–125.
[22] N.C. Wormald, *Analysis of greedy algorithms on graphs with bounded degrees, EuroComb* '01 *(Barcelona)*, Discrete Math., 273 (2003), pp. 235–260.
[23] N.C. Wormald, *Models of random regular graphs*, in Surveys in Combinatorics, 1999, London Math. Soc. Lecture Note Ser. 276, J.D. Lamb and D.A. Preece, eds., Cambridge University Press, Cambridge, 1999, pp. 239–298.
[24] N.C. Wormald, *The asymptotic connectivity of labelled regular graphs*, J. Combin. Theory Ser. B, 31 (1981), pp. 156–167.
[25] N.C. Wormald, *The differential equation method for random graph processes and greedy algorithms*, in Lectures on Approximation and Randomized Algorithms, M. Karoński and H.J. Prömel, eds., PWN, Warsaw, 1999, pp. 73–155.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 251–264

APPROXIMATING THE UNWEIGHTED k-SET COVER PROBLEM: GREEDY MEETS LOCAL SEARCH*

ASAF LEVIN†

Abstract. In the unweighted set cover problem we are given a set of elements $E = \{e_1, e_2, \ldots, e_n\}$ and a collection $\mathcal{F}$ of subsets of E. The problem is to compute a subcollection $SOL \subseteq \mathcal{F}$ such that $\bigcup_{S_j \in SOL} S_j = E$ and its size $|SOL|$ is minimized. When $|S| \leq k$ for all $S \in \mathcal{F}$, we obtain the unweighted k-set cover problem. It is well known that the greedy algorithm is an H_k-approximation algorithm for the unweighted k-set cover, where $H_k = \sum_{i=1}^{k} \frac{1}{i}$ is the kth harmonic number and that this bound on the approximation ratio of the greedy algorithm is tight for all constant values of k. Since the set cover problem is a fundamental problem, there is an ongoing research effort to improve this approximation ratio using modifications of the greedy algorithm. The previous best improvement of the greedy algorithm is an $(H_k - \frac{1}{2})$-approximation algorithm. In this paper we present a new $(H_k - \frac{196}{390})$-approximation algorithm for $k \geq 4$ that improves the previous best approximation ratio for all values of $k \geq 4$. Our algorithm is based on combining a local search during various stages of the greedy algorithm.

Key words. approximation algorithms, set cover

AMS subject classifications. 68Q25, 68W25, 68W40

DOI. 10.1137/060655225

1. Introduction.

In the WEIGHTED SET COVER PROBLEM we are given a set of elements $E = \{e_1, e_2, \ldots, e_n\}$ and a collection $\mathcal{F}$ of subsets of E, where $\cup_{S \in \mathcal{F}} S = E$ and each $S \in \mathcal{F}$ has a positive cost c_S. The goal is to compute a subcollection $SOL \subseteq \mathcal{F}$ such that $\bigcup_{S \in SOL} S = E$ and its cost $\sum_{S \in SOL} c_S$ is minimized. Such a subcollection of subsets is called a *cover*. When we consider instances of the WEIGHTED SET COVER such that each S_j has at most k elements ($|S| \leq k$ for all $S \in \mathcal{F}$), we obtain the WEIGHTED k-SET COVER PROBLEM. The UNWEIGHTED SET COVER PROBLEM and the UNWEIGHTED k-SET COVER PROBLEM are the special cases of the WEIGHTED SET COVER and of the WEIGHTED k-SET COVER, respectively, where $c_S = 1$ for all $S \in \mathcal{F}$.

It is well known (see [3]) that a greedy algorithm is an H_k-approximation algorithm for the weighted k-set cover, where $H_k = \sum_{i=1}^{k} \frac{1}{i}$ is the kth harmonic number and that this bound is tight even for the unweighted k-set cover problem (see, [13, 17]). For unbounded values of k, Slavík [21] showed that the approximation ratio of the greedy algorithm for the unweighted set cover problem is $\ln n - \ln \ln n + \Theta(1)$. Feige [6] proved that unless $NP \subseteq DTIME(n^{polylog\ n})$, the unweighted set cover problem cannot be approximated within a factor $(1 - \epsilon) \ln n$ for any $\epsilon > 0$. Raz and Safra [20] proved that if $P \neq NP$, then for some constant c, the unweighted set cover problem cannot be approximated within a factor $c \log n$. This result shows that the greedy algorithm is an asymptotically best possible approximation algorithm for the weighted and unweighted set cover problem (unless $NP \subseteq DTIME(n^{polylog\ n})$). The unweighted k-set cover problem is known to be NP-complete [14] and MAX SNP-hard for all $k \geq 3$ [4, 15, 18]. Another algorithm for the weighted set cover problem by Hochbaum [11] has an approximation ratio that depends on the maximum number of subsets that

*Received by the editors March 26, 2006; accepted for publication (in revised form) September 6, 2008; published electronically December 19, 2008.
http://www.siam.org/journals/sidma/23-1/65522.html
†Department of Statistics, The Hebrew University, 91905 Jerusalem, Israel (levinas@mscc.huji.ac.il).

contain any given element (the local-ratio algorithm of Bar-Yehuda and Even [2] has the same performance guarantee). See Paschos [19] for a survey on these results.

In spite of the above bad news, Goldschmidt, Hochbaum, and Yu [8] modified the greedy algorithm for the unweighted k-set cover and showed that the resulting algorithm has a performance guarantee of $H_k - \frac{1}{6}$. Halldórsson [9] presented an algorithm based on a local search that has an approximation ratio of $H_k - \frac{1}{3}$ for the unweighted k-set cover and a $(1.4 + \epsilon)$-approximation algorithm for the unweighted 3-set cover. Duh and Fürer [5] further improved this result and presented an $(H_k - \frac{1}{2})$-approximation algorithm for the unweighted k-set cover. We will base our algorithm on the algorithm of Duh and Fürer [5], and therefore we will review their algorithm and results in section 2.2. All of these improvements [8, 9, 5] are based on running the greedy algorithm until each new subset covers at most t new elements (where $t = 2$ in [8] and larger values of t in [9, 5]) and then switch to another algorithm.

Regarding approximation algorithms for the weighted k-set cover problem within a factor better than H_k, a first improvement step was given by Fujito and Okumura [7], who presented an $(H_k - \frac{1}{12})$-approximation algorithm for the k-set cover problem where the cost of each subset is either 1 or 2. More recently, Hassin and Levin [10] provided an $(H_k - \frac{k-1}{8k^9})$-approximation algorithm for the general weighted k-set cover problem.

The MAXIMUM SET PACKING PROBLEM is the following related problem: We are given a set of elements $E = \{e_1, e_2, \ldots, e_n\}$ and a collection $\mathcal{F}$ of subsets of E, where $\cup_{S \in \mathcal{F}} S = E$ and the goal is to compute a maximum size set packing, i.e., a subcollection $\mathcal{F}' \subseteq \mathcal{F}$ of disjoint subsets. The relation between the maximum set packing problem and the unweighted set cover problem is that the fractional version of the maximum set packing problem is the dual linear program of the fractional version of the unweighted set cover problem. Hurkens and Schrijver [12] proved that a local-search algorithm for the maximum set packing problem, where each subset in $\mathcal{F}$ has at most k elements, is a $(\frac{2}{k} - \epsilon)$-approximation algorithm. Therefore, this local-search algorithm has a better performance guarantee than the greedy selection rule that returns any maximal subcollection. The greedy selection rule has an approximation ratio of $\frac{1}{k}$.

Paper overview. In section 2 we review the greedy algorithm for the unweighted minimum k-set cover problem and its analysis, the semilocal optimization algorithm of [5], and then we present our improved algorithm. We analyze its performance in section 3, i.e., we show in Theorem 12 that our improved algorithm is an $(H_k - \frac{196}{390})$-approximation algorithm for the unweighted k-set cover problem, where $k \geq 4$, improving the earlier $(H_k - \frac{1}{2})$-approximation algorithm of [5]. We conclude in section 4 by discussing open questions.

2. Algorithms for the unweighted k-set cover problem.

In subsection 2.1 we review the greedy algorithm for the unweighted minimum k-set cover problem and its analysis. In subsection 2.2 we review the semilocal optimization algorithm of [5]. In subsection 2.3 we present our improved algorithm.

Given an input to the unweighted k-set cover problem, we let the *extended input* be defined over the same set of elements where the collection of subsets of the extended input is obtained from the input by including every subset of a subset in the input (i.e., the extended input is the closure of the input under taking subsets). We note that the extended input can be represented compactly by representing the maximal (under inclusion) subsets. A solution to the extended input is easily transformed into a solution for the original input by adding a superset which is included in the input

of each subset in the solution. This mapping can be maintained while creating the solution. For simplifying the presentation of the algorithms, we assume that they are solving the extended input. We also assume that the optimal solution is with respect to the extended input.

We start our study by stating a simplification lemma on the structure of the optimal solution.

LEMMA 1. *Without loss of generality, we may assume that the optimal solution to the (extended input of) a set cover instance satisfies that each element is covered by exactly one subset of the optimum.*

Proof. Let an optimal solution to the problem consist of a collection of sets S_j^*, $j \in J^*$, with $\cup_{j \in J^*} S_j^* = E$. We now construct another optimal solution formed of element-disjoint sets S_j', where $S_j' \subseteq S_j^*$ for all $j \in J^*$. To do that, we assign each element $e \in E$ to the smallest index set S_j^*, $j \in J^*$ that contains e, and, for all values of j, we let S_j' be the set of elements assigned to S_j^*. In the extended input the sets S_j' for all j belong to the collection $\mathcal{F}$, and the claim follows. ∎

We define a *j-set* to be a set with j elements. We fix an optimal solution OPT, and we say that a k-set is an *optimal k-set* if it is contained in OPT.

Given a partial cover $\mathcal{C}$ and an algorithm α, let $cost_\alpha(\mathcal{C})$ be the number of sets used by Algorithm α applied on the elements left uncovered by $\mathcal{C}$, and let $cost_{\alpha,1}(\mathcal{C})$ be the number of 1-sets among those.

2.1. The greedy algorithm. In this subsection we review the greedy algorithm for the unweighted k-set cover problem and the proof of its performance guarantee.

The greedy algorithm starts with an empty collection of subsets in the solution and no element being covered. Then, it iterates the following procedure until all elements are covered: Let w_S be the number of currently uncovered elements in a set $S \in \mathcal{F}$, and the current *ratio of* S is $r_S = \frac{1}{w_S}$. Let S^* be a set such that r_{S^*} is minimized. The algorithm adds S^* to the collection of subsets of the solution, defines the elements of S^* as covered, and assigns a *price* of r_{S^*} to all the elements that are now covered but were uncovered prior to this iteration (i.e., the elements that were first covered by S^*).

Johnson [13], Lovász [17], and Chvátal [3] showed that the greedy algorithm is an H_k-approximation algorithm for the unweighted k-set cover.

Chvátal's proof is the following: First, note that the cost of the greedy solution equals the sum of prices assigned to the elements. Second, consider a set S that belongs to an optimal solution OPT. Then, OPT pays 1 for S. Consider the elements of S in the order in which they are covered by the greedy algorithm breaking ties arbitrarily. When the algorithm covers the ith element of S, the algorithm could, instead, choose S as a feasible set with a current ratio of $\frac{1}{|S|-i+1}$. Therefore, the price assigned to the this element is at most $\frac{1}{|S|-i+1}$. It follows that the total price assigned to the elements of S is at most $\sum_{i=1}^{|S|} \frac{1}{|S|-i+1} = \sum_{i'=1}^{|S|} \frac{1}{i'} \leq H_k$, and therefore the approximation ratio of the greedy algorithm is at most H_k.

2.2. The semilocal optimization algorithm. Duh and Fürer [5] suggested the following procedure to approximate the unweighted 3-set cover problem. In a pure local improvement step, we replace a number of sets with fewer sets to form a new cover with a reduced cost. To define a semilocal step, they observed (see also [8]) that once the 3-sets are selected, the remaining instance can be solved optimally in polynomial time by reduction to maximum matching. Hence, to solve the unweighted 2-set cover instance results after selecting the 3-sets, they invoke the following Algorithm A.

ALGORITHM A FOR SOLVING OPTIMALLY UNWEIGHTED 2-SET COVER INSTANCE.

1. Find a maximum matching in the following graph: there is a vertex for each element, and an edge between two vertices if there is a 2-set consisting of this pair of elements.
2. Return the set of 2-sets corresponding to the edges of the maximum matching and the 1-sets of the uncovered elements (by the collection of 2-sets which we found).

Thus a local change in the 3-sets allows any global changes in the 2-sets and 1-sets, and such a change is called a *semilocal change*. They allowed the algorithm to remove one 3-set and insert at most a pair of 3-sets if one of the following happens: either the total cost is reduced, or the total cost remains the same and the number of 1-sets in the resulting solution is reduced (thus the total cost is the primary objective, whereas the number of 1-sets is a secondary objective). This results in the approximation algorithm (Algorithm B below) for the unweighted k-set cover of [5], which is useful mainly for $k = 3$.

ALGORITHM B FOR APPROXIMATING UNWEIGHTED k-SET COVER INSTANCE.

1. Greedily build a maximal collection $\mathcal{C}$ of disjoint sets, where each set in the collection contains at least three elements.
2. While there are sets $C \in \mathcal{C}$ and $C_1, C_2 \notin \mathcal{C}$ such that $\mathcal{C}' = (\mathcal{C} \setminus \{C\}) \cup \{C_1, C_2\}$ is a collection of disjoint sets, where each set in the collection contains at least three elements, and such that the following condition holds:
   $cost_A(\mathcal{C}') + |\mathcal{C}'| < cost_A(\mathcal{C}) + |\mathcal{C}|$ or $(cost_A(\mathcal{C}') + |\mathcal{C}'| = cost_A(\mathcal{C}) + |\mathcal{C}|$ and $cost_{A,1}(\mathcal{C}') < cost_{A,1}(\mathcal{C}))$.
   Replace $\mathcal{C}$ by $\mathcal{C}'$.
3. Apply Algorithm A on the remaining uncovered elements.

They showed that Algorithm B is a $\frac{4}{3}$-approximation algorithm for the unweighted 3-set cover problem. More precisely, the following proposition was proved in [5].

PROPOSITION 2. *Assume that an optimal solution for the unweighted* 3*-set cover instance has* b_1 1*-sets,* b_2 2*-sets, and* b_3 3*-sets. Then the solution that Algorithm* B *returns costs at most* $b_1 + b_2 + \frac{4}{3}b_3$ *(i.e.,* $cost_B(\emptyset) \le b_1 + b_2 + \frac{4}{3}b_3$*). Moreover, the number of* 1*-sets in the solution that the algorithm returns is at most* b_1 *(i.e.,* $cost_{B,1}(\emptyset) \le b_1$*).*

In order to extend their result to a better algorithm for larger values of k, they suggested the following Algorithm C.

ALGORITHM C FOR APPROXIMATING UNWEIGHTED k-SET COVER INSTANCE.

1. **Greedy Phase.** For $j = k$ down to 6 do:
   greedily choose a maximal collection of disjoint j-sets (each covering exactly j new elements).
2. **Restricted Phase.** For $j = 5$ down to 4 do:
   choose a maximal collection of disjoint j-sets (each covering exactly j new elements) with the restriction that the choice of these j-sets does not increase the number of 1-sets. That is, we add a j-set to the current collection of disjoint j-sets $\mathcal{C}$ and create a new collection of disjoint j-sets $\mathcal{C}'$ only if $cost_{B,1}(\mathcal{C}) \le cost_{B,1}(\mathcal{C}')$.
3. **Semilocal Optimization Phase.** Run the semilocal optimization algorithm (i.e., Algorithm B) on the remaining instance of the uncovered elements.

Duh and Fürer proved that this algorithm is an $(H_k - \frac{1}{2})$-approximation, and they also showed that this bound is tight for the semilocal optimization algorithm.

2.3. The improved algorithm. In this section we present our modification of the semilocal optimization algorithm where we use a local-search algorithm during the phase where each new set covers exactly four previously uncovered elements.

ALGORITHM D FOR APPROXIMATING UNWEIGHTED k-SET COVER INSTANCE—THE IMPROVED ALGORITHM.

1. **Greedy Phase.** For $j = k$ down to 6 do:
   greedily choose a maximal collection of disjoint j-sets (each covering exactly j new elements).
2. **Restricted Phase.** Choose a maximal collection of disjoint 5-sets (each covering exactly five new elements) with the restriction that the choice of these 5-sets does not increase the number of 1-sets. That is, we add a 5-set to the current collection of disjoint 5-sets $\mathcal{C}$ and create a new collection of disjoint 5-sets $\mathcal{C}'$ only if $cost_{B,1}(\mathcal{C}) \leq cost_{B,1}(\mathcal{C}')$.
3. **Restricted Local-Search Phase.**
   (a) Choose a maximal collection of disjoint 4-sets (each covering exactly four new elements) with the restriction that the choice of these 4-sets does not increase the number of 1-sets. That is, we add a 4-set to the current collection of disjoint 4-sets $\mathcal{C}$ and create a new collection of disjoint 4-sets $\mathcal{C}'$ only if $cost_{B,1}(\mathcal{C}) \leq cost_{B,1}(\mathcal{C}')$.
   (b) While there are 4-sets $C \in \mathcal{C}$ and $C_1, C_2 \notin \mathcal{C}$ such that $\mathcal{C}' = (\mathcal{C} \setminus \{C\}) \cup \{C_1, C_2\}$ is a collection of disjoint 4-sets and such that $cost_{B,1}(\mathcal{C}') \leq cost_{B,1}(\mathcal{C})$, replace $\mathcal{C}$ by $\mathcal{C}'$.
4. **Semilocal Optimization Phase.** Run the semilocal optimization algorithm (i.e., Algorithm B) on the remaining instance of the uncovered elements.

In Phase 3 we are using a local-search whose neighborhood is defined by removing one 4-set and inserting at least a pair of 4-sets as long as the number of 1-sets in the returned solution does not increase. The use of a local-search procedure is motivated by the approximation algorithm of [12] for the maximum set packing problem. That is, throughout the restricted phase (of Algorithm C), we try to maximize the number of sets in the collection of disjoint subsets that we add for a fixed value of the index j. Since a local search has proved to be superior heuristic for this task (with respect to its approximation ratio for this set packing problem), we suggest to replace the greedy construction for $j = 4$ in Algorithm C by the local-search approach. This improved phase is the cornerstone on which our improved approximation ratio is based.

To establish the time complexity of Algorithm D, we first note that Algorithm A is polynomial as it applies a maximum (cardinality) matching algorithm with time complexity $O(n^3)$. Hence, Algorithm B is also a polynomial time algorithm, as each iteration can be executed by trying all $O(m^3)$ triplets of sets and trying to increase the collection $\mathcal{C}$ with these sets. Such a test (for a given triplet of sets) is done in $O(n^3)$ by application of Algorithm A. Since the number of iterations of this loop of finding an increased collection of sets is bounded by $n/3$, the total time complexity of Algorithm B is $O(m^3n^4)$, that is, polynomial in the input length. Now consider Algorithm D. The time complexity of the greedy phase is $O(mn)$ per value of j and there are at most $k-5 < n$ such values, and hence the greedy phase takes $O(mn^2)$. Regarding the restricted phase, there are $O(m)$ sets to be considered, and each of them is tested by the application of Algorithm B, and hence this phase takes $O(m^4n^4)$. The restricted

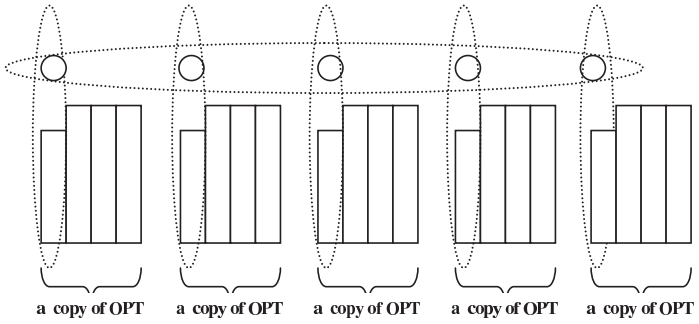


FIG. 1. *A demonstration of the instance I' for $k = 5$ in the proof of Lemma 4. Each circle represents a new element, and each dashed oval represent a new k-set, which is included only in I' and not in the copies of I.*

local-search phase is also polynomial as the number of 4-sets is $O(m)$, and each time we try to increase the number of 4-sets in $\mathcal{C}$, we try $O(m^3)$ triplet of 4-sets, and such a check is carried by running Algorithm B. Since the number of such iterations is bounded by $n/4$, we get time complexity of $O(m^6n^5)$ for this phase. The remaining part of the algorithm is a single execution of Algorithm B. Hence, the total time complexity of Algorithm D is $O(m^6n^5)$, that is, polynomial, and it returns a feasible solution. Therefore, we establish the following lemma.

LEMMA 3. *For every value of k, Algorithm* D *returns a feasible solution in polynomial time.*

In the next section we analyze the performance guarantee of Algorithm D.

3. The analysis of Algorithm D. In this section we analyze the performance guarantee of Algorithm D. We say that an element is an i-covered element if Algorithm D covers it by an i-set. We consider an optimal solution OPT and bound the performance guarantee of D. Recall that we assume that OPT is a partition of the element set E. We now further characterize the structure of OPT.

LEMMA 4. *If $k \geq 5$, then without loss of generality we can assume that each set of OPT is a k-set. If $k = 4$, then without loss of generality we can assume that each set of OPT is either a 3-set or a 4-set.*

Proof. Assume that the claim does not hold on an instance I. We create a new instance I' such that the optimal solution OPT' for I' costs k times the cost of OPT, and the solution returned by D on I' costs more than k times the solution returned by Algorithm D on I, and we will conclude that if there is a bad example for the algorithm, then there is a bad example for the algorithm that shows the same approximation ratio such that the property of the lemma holds.

To construct I' for $k \geq 5$, we first take k disjoint copies of the instance I. Then, we add new elements to the copies of the sets of OPT so that each set in this subcollection is a k-set. Note that the number of the new elements is divisible by k. Last, we add new disjoint k-sets covering these new elements. This is the new instance I' (see Figure 1 for an illustration).

Clearly, the optimal solution OPT' for I' is a union of k copies of OPT where we add the new elements to their corresponding set to make it a k-set. Hence, OPT' costs exactly k times the cost of OPT.

Now consider the execution of Algorithm D on the input I'. We can assume that the algorithm picks the new k-sets of the new elements in its first steps and then

continue like it acts on I on each of the k copies of I. Therefore, the cost of the solution returned by D on I' is strictly larger than k times the cost of the solution returned by D on I.

Thus the ratio $\frac{D(I')}{OPT'}$ is larger than the ratio $\frac{D(I)}{OPT}$, where $D(I')$ and $D(I)$ denote the cost of the solution returned by Algorithm D on instance I' and I, respectively. So the approximation ratio of Algorithm D can be computed by looking only at instances of the form of I', which satisfy the assumption of the lemma.

Now assume that $k = 4$. We apply a similar construction to the case of $k \geq 5$, with one difference. That is, we no longer add new elements to the copies of the 3-sets of OPT, and we make sure that each 4-set of new elements that we add has at most one new element from each set of OPT'. Once again, the cost of OPT' is exactly k times the cost of OPT. Now, the set of 4-sets of the new elements together with the copies of the original 4-sets returned by the Restricted Local-Search Phase on each copy of I, gives a feasible collection of 4-sets that cannot be extended. To see this last claim, note that, by deleting one 4-set of the new elements, none of the 4-sets which intersect it becomes disjoint to all other selected 4-sets. Hence, we can apply the same argument as in the case of $k \geq 5$. Therefore, the cost of the solution returned by D on I' is strictly larger than k times the cost of the solution returned by D on I.

Thus the ratio $\frac{D(I')}{OPT'}$ is larger than the ratio $\frac{D(I)}{OPT}$, where $D(I')$ and $D(I)$ denote the cost of the solution returned by Algorithm D on instance I' and I, respectively. So the approximation ratio of Algorithm D can be computed by looking only at instances of the form of I', which satisfy the assumption of the lemma. □

3.1. Sibling sets. We consider special 2-sets and 3-sets that are named *sibling sets* defined as follows (see [5] for introduction of this term): a sibling set is a 2-set or a 3-set S chosen by Algorithm D during the semilocal optimization phase, which intersects exactly two k-sets O_1, O_2 of OPT such that $|S \cap O_1| = 1$ and $S \cap O_1$ is the last element which is covered by Algorithm D. If this condition holds for both O_1 and O_2, this sibling set is called a *special sibling set*.

A sibling set is the result of the fact that the Semilocal Optimization Phase of Algorithm D does not create a new singleton, and therefore if an optimal k-set has $k-1$ covered elements at the end of Restricted Local-Search Phase of Algorithm D out of which at least one is either a 5-covered element or a 4-covered element, then the last element belongs to at least a 2-set (and is not a singleton).

The element of a sibling set S which is the last uncovered element of an optimal k-set, that is, the element of $S \cap O_1$, is called a *primary element*, and the other elements of S are called *secondary elements*. An element of a sibling set is called a *sibling element*. An element which is covered during phase 4 and is not a sibling element is called a *nonsibling element*.

LEMMA 5. *If a k-set S of OPT has a primary element, then all its elements which are covered during the Semilocal Optimization Phase are sibling elements.*

Proof. Assume that e is a primary element in S which belong to a sibling set S', and there is a nonsibling element in S which is covered during the Semilocal Optimization Phase. We note that Algorithm A could match e with its mates in S which are not sibling elements, without creating new singletons. Hence, the secondary elements of S' could be used during the Restricted Phase or the Local-Search Phase. Hence, S' is not a sibling set. □

3.2. Good and bad sets. We next partition the k-sets of OPT into bad sets and good sets. Let S be an optimal k-set. If $k \geq 5$, we say that S is a bad set if one of

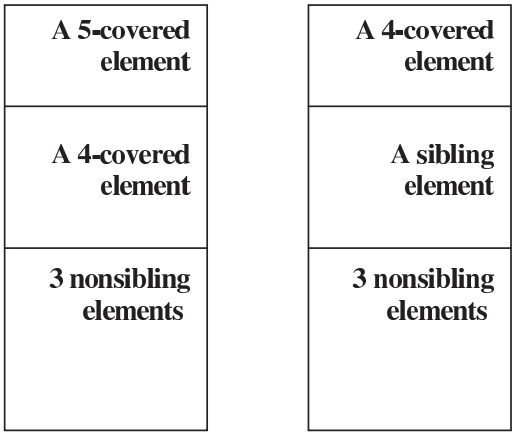


FIG. 2. *A demonstration of bad sets in the case* $k = 5$.

the following holds: Either at the end of Greedy Phase, S has exactly five uncovered elements from which exactly one element is 5-covered, exactly one is 4-covered, and none of the three remaining elements are sibling elements, or at the end of the Greedy Phase, S has exactly five uncovered elements from which none of the elements of S are 5-covered, exactly one 4-covered element, and exactly one element of S is a sibling element. We refer to Figure 2 for an illustration of this definition of a bad set.

If $k = 4$, then S is a bad set if exactly one of its elements is a 4-covered element and the other three elements are nonsibling elements. An optimal k-set that is not bad is a *good set*.

We next show that the proportion of good sets in OPT is not negligible. Denote by n_b the number of bad sets in OPT and by n_g the number of good sets in OPT.

LEMMA 6. $n_b \leq 12n_g$.

Proof. Consider a bad set S in OPT. At the beginning of phase 3, S has four uncovered elements such that none of these belong to a sibling set. Since S is a bad set, there is exactly one 4-covered element in S. Let S' be the set intersecting S, which is chosen by the algorithm in phase 3. If S' intersects only bad sets of OPT, then during phase 3 we could replace S' by the bad sets it intersects, and such a change is feasible because each such bad set has four elements that consist of a 4-set that we could add to the solution after the removal of S', without increasing the number of singletons. Hence, there is a good set $S'' \in OPT$ such that $S'' \cap S' \neq \emptyset$.

A good set $S \in OPT$ can intersect at most four sets that we choose during phase 3. These four sets can intersects at most 12 other sets of OPT. These 12 sets might be bad sets. Therefore, the claim follows. □

3.3. The pricing mechanism. Consider an element e, the price assigned to e which we denote by $price(e)$, is defined as follows.

- If e is an i-covered element where $i \geq 4$, then $price(e) = \frac{1}{i}$.
- If e is a member of a special sibling set, then $price(e) = \frac{1}{2}$.
- If e is a primary element of a sibling set, then $price(e) = \frac{4}{5}$, and if e is a secondary element, then $price(e) = \frac{1}{5}$.
- If e is a nonsibling element which is covered during phase 4, we assign its prices according to the value of $n(e)$, which denotes the number of nonsibling elements in the k-set of OPT which covers e:

- If $n(e) = 3$, then $price(e) = \frac{4}{9}$.
- If $n(e) = 2$, then $price(e) = \frac{1}{2}$.
- If $n(e) = 1$ and at the end of the Greedy Phase there are at least two uncovered elements in the optimal k-set which covers e, then $price(e) = \frac{1}{2}$.
- Otherwise, that is, if $price(e)$ is not already set by the previous cases, then $price(e) = 1$.

Note that if $n(e) = 1$ and at the end of the Greedy Phase there are at least two uncovered elements in the optimal k-set which covers e, then the other uncovered element at the end of the Greedy Phase is not a primary element of a nonspecial sibling set.

LEMMA 7. *The cost of the solution returned by Algorithm* D *is at most the total price of all the elements.*

Proof. We clearly assigned a total of a unit price for each selected set in phases 1, 2, and 3, and for sibling sets that the algorithm selects.

As for the other sets, we denote by $b_3(OPT)$ the number of k-sets of OPT, with exactly three nonsibling elements, and we denote by $b_2(OPT)$ the number of k-sets of OPT, with exactly two nonsibling elements. By Proposition 2, the number of the nonsibling sets that the algorithm selects during phase 4 is at most $\frac{4}{3}b_3(OPT) + b_2(OPT)$, that is, the total price of the nonsibling elements. □

3.4. Bounding the total price assigned to the elements of an optimal k-set. For a set of items S, we denote by $price(S)$ the total price assigned to the elements of S.

LEMMA 8. *Assume that $k \geq 4$. Let S be an optimal bad k-set. Then, $price(S) \leq \rho_b = H_k - \frac{1}{2}$.*

Proof. If $k \geq 5$, then the jth covered element from S during the Greedy Phase is assigned a price of at most $\frac{1}{k-j+1}$, the 5-covered element is assigned a price of $\frac{1}{5}$ (if it exists), the sibling element is assigned a price of $\frac{1}{5}$ (if it exists), the 4-covered element is assigned a price of $\frac{1}{4}$, and each of the remaining three elements is assigned a price of $\frac{4}{9}$. Hence, $price(S) \leq \sum_{i=6}^{k} \frac{1}{i} + \frac{1}{5} + \frac{1}{4} + 3\frac{4}{9} = H_k - \frac{1}{2} = \rho_b$. If $k = 4$, then S has a single 4-covered element that pays a price of $\frac{1}{4}$, and each of the three remaining elements is assigned a price of $\frac{4}{9}$. So again $price(S) = H_k - \frac{1}{2} = \rho_b$. □

Before bounding the total price assigned to an optimal good k-set, we bound the total price of the items covered during the Semilocal Optimization Phase of an optimal k-set. These bounds will be used later in the upper bound proof of the total price assigned to an optimal good k-set. We denote by N_g the number of the elements of S that remain uncovered at the end of Greedy Phase. Note that $N_g \leq 5$. We denote by N_r (N_l) the number of the elements of S that are covered during the Restricted Phase (the Restricted Local-search Phase). We denote by N_s the number of sibling elements of S that are covered during the Semilocal Optimization Phase. We denote by N_n the number of nonsibling elements of S that are covered during the Semilocal Optimization Phase. Then, $N_s + N_n = N_g - (N_r + N_l)$ is the number of elements of S which are covered during the Semilocal Optimization Phase. Let S' be the subset of S consisting of the elements covered during the Semilocal Optimization Phase. The following lemma bound $price(S')$ as a function of $N_s + N_n = |S'|$.

LEMMA 9.

1. *If $N_s + N_n = 5$, then $price(S') \leq \frac{26}{15}$.*
2. *If $N_s + N_n = 4$, then $price(S') \leq \frac{23}{15}$.*
3. *If $N_s + N_n = 3$, then $price(S') \leq \frac{4}{3}$.*
4. *If $N_s + N_n = 2$, then $price(S') \leq 1$.*

5. *If* $N_s + N_n = 1$ *and* $N_g \geq 2$*, then* $price(S') \leq \frac{4}{5}$.
6. *If* $N_s + N_n = 1$ *and* $N_g = 1$*, then* $price(S') \leq 1$.

Proof. Assume that $N_s + N_n = 5$. Then, since S' and each of its 4-subsets are candidates to be added to the collection of disjoint 4-sets during the Restricted Local-Search Phase and we choose not to add them, we conclude that at least two elements of S' are sibling elements, i.e., $N_s \geq 2$. If one of the elements of S' is a primary element, then all other elements in S' are secondary elements, and $price(S') = \frac{4}{5}+\frac{1}{5}+\frac{1}{5}+\frac{1}{5}+\frac{1}{5} < \frac{26}{15}$. Otherwise, all sibling elements of S' are secondary elements, and each of these pays $\frac{1}{5}$.

- If $N_s = 2$, then each of the nonsibling element of S' pays $\frac{4}{9}$, and hence $price(S') = 3 \cdot \frac{4}{9} + 2 \cdot \frac{1}{5} = \frac{26}{15}$.
- If $N_s = 3$, then each of the nonsibling element of S' pays $\frac{1}{2}$, and hence $price(S') = 2 \cdot \frac{1}{2} + 3 \cdot \frac{1}{5} < \frac{26}{15}$.
- If $N_s = 4$, then the unique nonsibling element of S' pays $\frac{1}{2}$, and hence $price(S') = 1 \cdot \frac{1}{2} + 4 \cdot \frac{1}{5} < \frac{26}{15}$.
- If $N_s = 5$, then $price(S') = 5 \cdot \frac{1}{5} < \frac{26}{15}$.

This completes the proof of part 1 of Lemma 9.

Next assume that $N_s + N_n = 4$. Then, since S' is a candidate to be added to the collection of disjoint 4-sets during the Restricted Local-Search Phase and we choose not to add it, we conclude that at least one element of S' is a sibling element, i.e., $N_s \geq 1$. If one of the elements of S' is a primary element, then all other elements in S' are secondary elements, and $price(S') = \frac{4}{5} + 3 \cdot \frac{1}{5} < \frac{23}{15}$. Otherwise, all sibling elements of S' are secondary elements, and each of these pays $\frac{1}{5}$.

- If $N_s = 1$, then each of the nonsibling element of S' pays $\frac{4}{9}$, and hence $price(S') = 3 \cdot \frac{4}{9} + 1 \cdot \frac{1}{5} = \frac{23}{15}$.
- If $N_s = 2$, then each of the nonsibling element of S' pays $\frac{1}{2}$, and hence $price(S') = 2 \cdot \frac{1}{2} + 2 \cdot \frac{1}{5} < \frac{23}{15}$.
- If $N_s = 3$, then the unique nonsibling element of S' pays $\frac{1}{2}$, and hence $price(S') = 1 \cdot \frac{1}{2} + 3 \cdot \frac{1}{5} < \frac{23}{15}$.
- If $N_s = 4$, then $price(S') = 4 \cdot \frac{1}{5} < \frac{23}{15}$.

This completes the proof of part 2 of Lemma 9.

Next assume that $N_s + N_n = 3$. If one of the elements of S' is a primary element, then all other elements in S' are secondary elements, and $price(S') = \frac{4}{5} + 2 \cdot \frac{1}{5} < \frac{4}{3}$. Otherwise, all sibling elements of S' are secondary elements, and each of these pays $\frac{1}{5}$.

- If $N_s = 0$, then each element of S' pays $\frac{4}{9}$, and hence $price(S') = 3 \cdot \frac{4}{9} = \frac{4}{3}$.
- If $N_s = 1$, then each of the nonsibling element of S' pays $\frac{1}{2}$, and hence $price(S') = 2 \cdot \frac{1}{2} + \frac{1}{5} < \frac{4}{3}$.
- If $N_s = 2$, then the unique nonsibling element of S' pays $\frac{1}{2}$, and hence $price(S') = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{5} < \frac{4}{3}$.
- If $N_s = 3$, then $price(S') = 3 \cdot \frac{1}{5} < \frac{4}{3}$.

This completes the proof of part 3 of Lemma 9.

Next assume that $N_s + N_n = 2$. If one of the elements of S' is a primary element, then the other element in S' is a secondary element, and $price(S') = \frac{4}{5} + \frac{1}{5} = 1$. Otherwise, all sibling elements of S' are secondary elements, and each of these pays $\frac{1}{5}$.

- If $N_s = 0$, then each element of S' pays $\frac{1}{2}$, and hence $price(S') = 2 \cdot \frac{1}{2} = 1$.
- If $N_s = 1$, then the unique nonsibling element of S' pays $\frac{1}{2}$, and hence $price(S') = 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{5} < 1$.
- If $N_s = 2$, then $price(S') = 2 \cdot \frac{1}{5} < 1$.

This completes the proof of part 4 of Lemma 9.

Finally, we assume that $N_s + N_n = 1$. If $N_g \geq 2$, then we did not assign this element a unit price, and hence we assign it at most $\frac{4}{5}$, which is the maximum price of an element excluding one. If $N_g = 1$, the claim is trivial as every element is assigned at most a unit of price. This completes the proof of parts 5 and 6 of Lemma 9. ∎

LEMMA 10. *If $N_g = 5$ and $N_r \leq 1$, then S has an element that pays exactly $\frac{1}{5}$.*

Proof. By the maximality of the sets that we choose during the Restricted Phase, we conclude that if $N_r = 0$, then S has a secondary element. In both cases, S has an element that pays $\frac{1}{5}$. ∎

LEMMA 11. *Assume that $k \geq 4$. Let S be an optimal good k-set. Then, $price(S) \leq \rho_g = H_k - \frac{16}{30}$.*

Proof. Our proof is based on a detailed case analysis. These cases are according to the values of k (either four or at least five), N_g, N_r, N_l, and $N_s + N_n$.

First assume that $k = 4$. Then, the Greedy Phase and the Restricted Phase do not select sets, and therefore $N_g = 4$ and $N_r = 0$.

- Assume that $N_l = 4$. Then, each element of S is covered during the Restricted Local-Search Phase and pays a price of $\frac{1}{4}$. Therefore, $price(S) = 1 < H_4 - \frac{16}{30} = \rho_g$.
- Assume that $N_l = 3$. Then, each element of S which is covered during Restricted Local-Search Phase pays a price of $\frac{1}{4}$, and, by Lemma 9, the remaining element pays a price of at most $\frac{4}{5}$. Therefore, $price(S) \leq \frac{3}{4} + \frac{4}{5} = \frac{93}{60} = \frac{125}{60} - \frac{32}{60} = H_4 - \frac{16}{30} = \rho_g$.
- Assume that $N_l = 2$. Then, each element of S which is covered during Restricted Local-Search Phase pays a price of $\frac{1}{4}$. By Lemma 9, the two remaining elements pay a total price of at most 1. Thus, $price(S) \leq \frac{3}{2} < \rho_g$.
- Assume that $N_l = 1$. Then, the element of S which is covered during Restricted Local-Search Phase pays a price of $\frac{1}{4}$. Since S is a good set, it contains at least one element that belongs to a sibling set that pays $\frac{1}{5}$ (since $N_l = 1$, it is not the primary element). The other two elements of S have a total price of at most $\max\{2 \cdot \frac{1}{2}, 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{5}, \frac{4}{5} + \frac{1}{5}, 2 \cdot \frac{1}{5}\} = 1$ (the arguments of the maximum are according to the number of sibling elements). Therefore, $price(S) \leq \frac{1}{4} + \frac{1}{5} + 1 = \frac{87}{60} < \frac{93}{60} = \rho_g$.
- Assume that $N_l = 0$. By Lemma 9, $price(S) \leq \frac{23}{15} < \rho_g$.

It remains to consider the case where $k \geq 5$. First note that, by the greedy selection rule during the greedy phase, we conclude that $N_g \leq 5$. Moreover, the jth covered element from S during the greedy phase (for $1 \leq j \leq k-5$) is assigned a price of at most $\frac{1}{k-j+1}$. So, the first $k-5$ elements which are covered by the algorithm pay a total price of at most $H_k - H_5$.

- Assume that $N_g \leq 2$. Then, the $k-4$th, the $k-3$rd, and the $k-2$nd covered elements from S are covered during the Greedy Phase and therefore assigned a price of at most $\frac{1}{6}$ for each. The last two elements of S are assigned a total price of at most $\max\{2 \cdot \frac{1}{4}, \frac{4}{5} + \frac{1}{4}, 1\} = \frac{21}{20}$ (the arguments of the maximum are according to the value of N_l). Therefore, $price(S) \leq H_k - H_5 + \frac{3}{6} + \frac{21}{20} = H_k - H_5 + \frac{31}{20} = H_k - \frac{137}{60} + \frac{93}{60} = H_k - \frac{44}{60} < \rho_g$.
- Assume that $N_g = 3$. Then, the $k-4$th and the $k-3$rd covered elements from S are covered during the Greedy Phase and therefore assigned a price of at most $\frac{1}{6}$ for each.
  - If $N_r + N_l = 0$, then the last three elements of S are covered during the Semilocal Optimization Phase and, by Lemma 9, pay a total price of at most $\frac{4}{3}$. Therefore, $price(S) \leq H_k - H_5 + \frac{2}{6} + \frac{4}{3} = H_k - \frac{137}{60} + \frac{5}{3} =$

$H_k - \frac{37}{60} < \rho_g$. Note that in the remaining cases (of $N_r + N_l$), it suffices to show that the last three elements of S pay a total price of at most $\frac{4}{3}$.

- If $N_r + N_l = 1$, then the last two elements of S are covered during the Semilocal Optimization Phase and, by Lemma 9, pay a total price of at most 1. The $k-2$nd element of S is covered during either the Restricted Phase or the Restricted Local-Search Phase, and so it pays a price of at most $\frac{1}{4}$. Therefore, the last three elements of S pay a total price of at most $\frac{5}{4} < \frac{4}{3}$, and again $price(S) < \rho_g$.
- If $N_r + N_l = 2$, then, by Lemma 9, the last uncovered element pays at most $\frac{4}{5}$. The $k-2$nd and the $k-1$st covered elements from S are covered during either the Restricted Phase or the Restricted Local-Search Phase, and therefore each of these is assigned a price of at most $\frac{1}{4}$. Again, the last three elements of S pay at most $\frac{4}{5} + \frac{2}{4} < \frac{4}{3}$, and therefore $price(S) < \rho_g$.
- If $N_r + N_l = 3$, then each of the last three elements of S pays a price of at most $\frac{1}{4}$, and, in total, they pay less than $\frac{4}{3}$. Therefore, $price(S) < \rho_g$.

- Assume that $N_g = 4$. Then, the $k-4$th covered element from S is covered during the Greedy Phase and therefore pays a price of at most $\frac{1}{6}$, and the set of elements from S that are covered during the Greedy Phase pays a total price of at most $H_k - H_5 + \frac{1}{6}$. By Lemma 9, if $N_l = N_r = 0$, then $price(S) \le H_k - H_5 + \frac{1}{6} + \frac{23}{15} = H_k - \frac{137}{60} + \frac{102}{60} < \rho_g$. Otherwise, there is at least one element which is covered during the Restricted Phase or the Restricted Local-Search Phase, and hence it pays at most $\frac{1}{4}$. The other three elements pay a total price of at most $\max\{\frac{4}{3}, 1 + \frac{1}{4}, \frac{4}{5} + \frac{2}{4}, \frac{3}{4}\} = \frac{4}{3}$ (the arguments of the maximum are according to the value of N_l). Therefore, $price(S) \le H_k - H_5 + \frac{1}{6} + \frac{1}{4} + \frac{4}{3} = H_k - \frac{137}{60} + \frac{105}{60} = H_k - \frac{32}{60} = \rho_g$.
- Assume that $N_g = 5$. Then, the set of elements from S that are covered during the Greedy Phase pays a total price of at most $H_k - H_5$. Each of the elements of S that is covered during Phase 3 pays a price of $\frac{1}{4}$.
  - Assume that $N_r = N_l = 0$. By Lemma 9, $price(S) \le H_k - H_5 + \frac{26}{15} = H_k - \frac{137}{60} + \frac{104}{60} = H_k - \frac{33}{60} < \rho_g$.
  - Assume that $N_r = 1$ and $N_l = 0$. The element of S that is covered during the Restricted Phase pays a price of $\frac{1}{5}$. By Lemma 9, $price(S) \le H_k - H_5 + \frac{1}{5} + \frac{23}{15} = H_k - \frac{33}{60} < \rho_g$.
  - Assume that $N_r \ge 2$. The elements of S that are covered during the Restricted Phase pay a price of $\frac{1}{5}$ each. The last three elements pay a total price of at most $\max\{\frac{4}{3}, \frac{1}{4} + 1, 2 \cdot \frac{1}{4} + \frac{4}{5}, 3 \cdot \frac{1}{4}\} = \frac{4}{3}$ (the arguments of the maximum are according to the value of $N_s + N_n$). Therefore, $price(S) \le H_k - H_5 + \frac{2}{5} + \frac{4}{3} = H_k - \frac{33}{60} < \rho_g$.
  - Assume that $N_r \le 1$ and $N_l = 1$. Since S is a good set, we conclude that either $N_r = 1$ and S has an element that belongs to a sibling set, or S has at least two elements that belong to sibling sets. The element of S that is covered during the Restricted Phase (if it exists) pays a price of $\frac{1}{5}$, the element of S that is covered during the Restricted Local-Search Phase pays a price of $\frac{1}{4}$, and each secondary element of S pays $\frac{1}{5}$. The two last remaining elements have a total price of at most 1. Therefore, $price(S) \le H_k - H_5 + \frac{1}{5} + \frac{1}{4} + \frac{1}{5} + 1 = H_k - \frac{137}{60} + \frac{99}{60} = H_k - \frac{38}{60} < \rho_g$.
  - Assume that $N_r \le 1$ and $N_l = 2$. By Lemma 10, S has an element that pays $\frac{1}{5}$. The two remaining elements which are covered during the Semilocal Optimization Phase pay a total price of at most 1. Therefore,

$price(S) \leq H_k - H_5 + \frac{1}{5} + \frac{2}{4} + 1 = H_k - \frac{137}{60} + \frac{102}{60} = H_k - \frac{35}{60} < \rho_g$.

- Assume that $N_r \leq 1$ and $N_l = 3$. By Lemma 10, S has an element that pays $\frac{1}{5}$. By Lemma 9, the element which is covered during the Semilocal Optimization Phase pays at most $\frac{4}{5}$. Therefore, $price(S) \leq H_k - H_5 + \frac{1}{5} + \frac{3}{4} + \frac{4}{5} = H_k - \frac{137}{60} + \frac{105}{60} = H_k - \frac{32}{60} = \rho_g$.
- Assume that $N_r \leq 1$ and $N_l = 4$. By Lemma 10, S has an element that pays $\frac{1}{5}$. Therefore, $price(S) \leq H_k - H_5 + \frac{1}{5} + \frac{4}{4} = H_k - \frac{137}{60} + \frac{72}{60} = H_k - \frac{65}{60} < \rho_g$. □

3.5. Proving the approximation ratio of Algorithm D.

THEOREM 12. *Algorithm D is an* $(H_k - \frac{196}{390})$*-approximation algorithm for the unweighted k-set cover problem.*

Proof. By Lemma 3, the algorithm returns a feasible solution in polynomial time. It remains to establish its approximation ratio.

$$\begin{aligned} D &\leq n_g \cdot \rho_g + n_b \cdot \rho_b \\ &= n_g \cdot \left(H_k - \frac{16}{30}\right) + n_b \cdot \left(H_k - \frac{1}{2}\right) \\ &\leq (n_g + n_b) \cdot \left[\frac{1}{13} \cdot \left(H_k - \frac{16}{30}\right) + \frac{12}{13} \cdot \left(H_k - \frac{1}{2}\right)\right] \\ &= OPT \cdot \left[\frac{1}{13} \cdot \left(H_k - \frac{16}{30}\right) + \frac{12}{13} \cdot \left(H_k - \frac{1}{2}\right)\right] \\ &= OPT \cdot \left(H_k - \frac{196}{390}\right), \end{aligned}$$

where the first inequality follows by Lemma 7, the first equation follows by Lemma 8 and Lemma 11, the second inequality follows by Lemma 6, the second equation follows because the cost of OPT is exactly $n_b + n_g$, and the last equation follows by simple algebra. □

4. Concluding remarks. In this paper we addressed the fundamental problem of the unweighted k-set cover problem and introduced an improvement over the previously best known algorithm for all values of k such that $k \geq 4$. Although we obtain a small improvement over the algorithm of Duh and Fürer [5], we think that our analysis is not tight, and the approximation ratio of our algorithm can be improved. Improving the analysis of our Algorithm D is left for future research.

In this paper we showed that incorporating a local-search procedure in various stages of the greedy algorithm, instead of only where each set has at most three uncovered elements, provides a better approximation ratio. We conjecture that incorporating local-search procedures in each greedy phase decreases the approximation ratio further. Such an algorithm replaces the Greedy phase by the following phase.

Improved Phase. For $j = k, k-1, k-2, \ldots, 6$ do: apply local-search to choose an approximated maximum size collection of j-sets (each covering exactly j new elements).

It is easily noted that using the Improved Phase instead of the Greedy Phase in Algorithm D does not harm the approximation ratio of the resulting algorithm. We leave the analysis of this improved algorithm for future research. Following an extended abstract version of this paper [16], Athanassopoulos, Caragiannis, and Kak-

lamanis [1] showed that this improved step indeed improves the approximation ratio of the resulting algorithm.

REFERENCES

[1] S. Athanassopoulos, I. Caragiannis, and C. Kaklamanis, *Analysis of approximation algorithms for k-set cover using factor-revealing linear programs,* in Proceedings of FCT 2007, Budapest, Hungary, 2007, pp. 52–63.

[2] R. Bar-Yehuda and S. Even, *A linear time approximation algorithm for the weighted vertex cover problem,* J. Algorithms, 2 (1981), pp. 198–203.

[3] V. Chvátal, *A greedy heuristic for the set-covering problem,* Math. Oper. Res., 4 (1979), pp. 233–235.

[4] P. Crescenzi and V. Kann, *A Compendium of NP Optimization Problems,* http://www.nada.kth.se/theory/problemlist.html.

[5] R. Duh and M. Fürer, *Approximation of k-set cover by semi local optimization,* in Proceedings of the 29th Annual ACM STOC 1997, El Paso, TX, 1997, pp. 256–264.

[6] U. Feige, *A threshold of* $\ln n$ *for approximating set cover,* J. ACM, 45 (1998), pp. 634–652.

[7] T. Fujito and T. Okumura, *A modified greedy algorithm for the set cover problem with weights* 1 *and* 2*,* in Proceedings of ISAAC 2001, Christchurch, New Zealand, 2001, pp. 670–681.

[8] O. Goldschmidt, D. S. Hochbaum, and G. Yu, *A modified greedy heuristic for the set covering problem with improved worst case bound,* Inform. Process. Lett., 48 (1993), pp. 305–310.

[9] M. M. Halldórsson, *Approximating k set cover and complementary graph coloring,* in Proceedings of IPCO 1996, Vancouver, BC, 1996, pp. 118–131.

[10] R. Hassin and A. Levin, *A better-than-greedy approximation algorithm for the minimum set cover problem,* SIAM J. Comput., 35 (2005), pp. 189–200.

[11] D. S. Hochbaum, *Approximation algorithms for the set covering and vertex cover problems,* SIAM J. Comput., 11 (1982), pp. 555–556.

[12] C. A. J. Hurkens and A. Schrijver, *On the size of systems of sets every t of which have an SDR, with an application to the worst-case ratio of heuristics for packing problems,* SIAM J. Discrete Math., 2 (1989), pp. 68–72.

[13] D. S. Johnson, *Approximation algorithms for combinatorial problems,* J. Comput. System Sci., 9 (1974), pp. 256–278.

[14] R. M. Karp, *Reducibility among combinatorial problems,* in Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, eds., Plenum Press, New York, 1972, pp. 85–103.

[15] S. Khanna, R. Motwani, M. Sudan, and U. V. Vazirani, *On syntactic versus computational views of approximability,* SIAM J. Comput., 28 (1998), pp. 164–191.

[16] A. Levin, *Approximating the unweighted k-set cover problem: Greedy meets local search,* in Proceedings of WAOA 2006, Zurich, 2006, pp. 290–301.

[17] L. Lovász, *On the ratio of optimal integral and fractional covers,* Discrete Math., 13 (1975), pp. 383–390.

[18] C. H. Papadimitriou and M. Yannakakis, *Optimization, approximation and complexity classes,* J. Comput. System Sci., 43 (1991), pp. 425–440.

[19] V. T. Paschos, *A survey of approximately optimal solutions to some covering and packing problems,* ACM Comput. Surveys, 29 (1997), pp. 171–209.

[20] R. Raz and S. Safra, *A sub-constant error-probability low-degree test, and sub-constant error-probability PCP characterization of NP,* in Proceedings of the 29th Annual ACM STOC 1997, El Paso, TX, 1997, pp. 475–484.

[21] P. Slavík, *A tight analysis of the greedy algorithm for set cover,* J. Algorithms, 25 (1997), pp. 237–254.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 265–277

ASYMPTOTIC BOUNDS ON THE INTEGRITY OF GRAPHS AND SEPARATOR THEOREMS FOR GRAPHS*

D. BENKO†, C. ERNST‡, AND D. LANPHIER‡

Abstract. In this paper we study the integrity of certain graph families. These include planar graphs, graphs with a given genus, graphs on the d-dimensional integer lattice $\mathbb{Z}^d$, and graphs that have no K_h-minor. We give upper bounds for the integrity in terms of the order n of the graph. We also give lower bounds for box-graphs in $\mathbb{Z}^d$. As a consequence, the integrity of planar graphs is on the order of $n^{2/3}$, where 2/3 is the best possible exponent.

Key words. integrity, planar graphs, lattice graphs, separators

AMS subject classification. Primary, 57M25

DOI. 10.1137/070692698

1. Introduction.

The *integrity* of a finite graph G is

$$I(G) = \min_{S \subset V(G)} \big(|S| + \tau(G \setminus S)\big),$$

where $\tau(G \setminus S)$ denotes the size of the largest component of $G \setminus S$. The integrity can be thought of as a measurement of connectivity of a graph. $|S|$ measures the amount of work needed to damage or disconnect a graph, while $\tau(G \setminus S)$ is a measure of how much of the graph is still intact. The integrity is the sum of these two quantities and was first introduced by Barefoot, Entringer, and Swart [4] inspired by the idea to measure a computer network's vulnerability.

Throughout this paper we assume that G is a graph with n vertices. It is easy to see that for the complete graph K_n, we have $I(K_n) = n$, and there are examples of simple, regular graphs with integrity of the order of n^α for any $0 \leq \alpha \leq 1$. However, the exact integrity of a given graph is difficult to compute. In fact, only for very simple graph families is the exact integrity known, so even establishing upper bounds for the integrity of large graph families is a worthwhile goal. See [3] and [8] for further information about the integrity of graphs.

A graph H is a *minor* of a graph G if H can be obtained from a subgraph of G by contracting edges. An *H-minor* of G is a minor of G isomorphic to H. The *genus* g of a graph G is the smallest genus of all surfaces (compact orientable 2-manifolds) on which G can be properly embedded. In this paper we show that the integrity of graphs with no K_h-minor is $O(n^{2/3})$, where $h \geq 3$ is fixed. We give explicit upper bounds with particular attention to the case of planar graphs. The key property is that such graphs possess separator theorems of the form found in [1, 2, 5, 6, 7, 9]; see also section 2.

Our main results are Theorems 1.1, 1.2, 1.3, and 1.4 below.

*Received by the editors May 22, 2007; accepted for publication (in revised form) May 17, 2008; published electronically January 7, 2009. This work is partially supported by NSF grant DMS-0712997.

http://www.siam.org/journals/sidma/23-1/69269.html

†University of South Alabama, Mobile, AL 36617 (dbenko@jaguar1.usouthal.edu).

‡Department of Mathematics, Western Kentucky University, Bowling Green, KY 42101 (claus.ernst@wku.edu, dominic.lanphier@wku.edu).

THEOREM 1.1. *Let G be a graph of order n with no K_h-minor for fixed $h \geq 3$. Then for $n \geq 119h^3$,*

$$I(G) \leq 10.9hn^{2/3} - 13.1h^{3/2}n^{1/2}.$$

THEOREM 1.2. *Let G be a planar graph of order n. Then for $n \geq 535$,*

$$I(G) \leq 18n^{2/3} - 27.9n^{1/2}.$$

THEOREM 1.3. *Let G be a graph of genus g and of order n. Then for $n \geq 713(2g+1)$,*

$$I(G) \leq 19.8(2g+1)^{1/3}n^{2/3} - 32.2(2g+1)^{1/2}n^{1/2}.$$

It follows that

$$I(G) = O\left(n^{2/3}\right)$$

for the graph families in Theorems 1.1, 1.2, and 1.3. Theorem 1.5 below shows that for planar graphs, 2/3 is the best possible exponent.

Let $\mathbb{Z}^d$ denote the lattice graph where vertices are the points in $\mathbb{R}^n$ with integer coordinates, and vertices are adjacent if and only if their Euclidean distance is 1. A subgraph of $\mathbb{Z}^d$ which forms a rectangular box whose sides are parallel to the axes will be called a *box-graph*. The dimensions of a box-graph are the number of vertices lying on the edges of the box. (So, each dimension is the length of an edge plus 1.) The order of a box-graph is the product of its dimensions.

The theorem below provides a formula for calculating the integrity of a box-graph up to a constant factor depending on the dimension d only.

THEOREM 1.4. *Let G be a box-graph in $\mathbb{Z}^d$ with dimensions $a_1, \ldots, a_d$, where $a_1 \geq \cdots \geq a_d$ and set $m(G) = \sqrt[2]{a_1} + \sqrt[3]{a_1a_2} + \cdots + \sqrt[d+1]{a_1a_2 \ldots a_d}$. Then*

$$c_d \leq I(G) \Big/ \left(\frac{|V(G)|}{m(G)}\right) \leq C_d, \tag{1.1}$$

where the constants c_d and C_d depend on d only.

Theorem 1.4 is equivalent to Lemma 4.2 in section 4 (the constants may differ). For planar box-graphs, the proof of Lemma 4.2 gives the following result.

THEOREM 1.5. *Let G be a planar box-graph of order n with dimensions a_1, a_2, and $a_1 \geq a_2$ (so $n = a_1a_2$). If $a_2 \geq 2\sqrt{a_1}$, then*

$$0.00136n^{2/3} \leq I(G) \leq 5.22n^{2/3}.$$

If $a_2 < 2\sqrt{a_1}$, then

$$0.00136\sqrt{a_1}a_2 \leq I(G) \leq 5.22\sqrt{a_1}a_2.$$

Another special case of Theorem 1.4 is the following.

THEOREM 1.6. *Let G be the box-graph in $\mathbb{Z}^d$ which forms a cube. Let "a" denote the dimensions of the cube, so G has order $n = a^d$. Then there exist constants c_d and C_d depending on d alone such that*

$$c_d n^{\frac{d}{d+1}} \leq I(G) \leq C_d n^{\frac{d}{d+1}}.$$

Theorem 1.4 is also demonstrated in the following example for "flat" prism box-graphs.

Example 1.7. Let $G \subset \mathbb{Z}^3$ be a box-graph with dimensions a_1, a_2, a_3 and assume that $b := a_1 = a_2 \geq a_3 =: a$. In the notation of Lemma 4.2, $A_0 = 1$, $A_1 = \sqrt{b}$, $A_2 = \sqrt[3]{b^2}$, and $A_3 = \sqrt[4]{b^2 a}$. If $b \geq 4$ and $a < 2\sqrt[3]{b^2}$, then, in Lemma 4.2, we have $N = 2$. Now $|V(G)|/A_2 = ab^2/\sqrt[3]{b^2} = ab^{4/3}$, and so $c_d^* ab^{4/3} \leq I(G) \leq C_d^* ab^{4/3}$.

2. Separator theorems. For $A \subset V(G)$, we denote by $G[A]$ the induced subgraph of G. (This is a graph whose vertex set is A and where two vertices in $G[A]$ are connected by an edge if and only if they are connected by an edge in the graph G.)

PROPOSITION 2.1. *Let $0 < \alpha \leq 1$ and $1 \leq c$. Let $\mathcal{G}_\alpha(c)$ be a family of graphs so that for any $G \in \mathcal{G}_\alpha(c)$, there exists a vertex partition $V(G) = A \cup B \cup C$, where $|A|, |B| \leq (2/3)|V(G)|$, $|C| \leq c|V(G)|^\alpha$, and no vertex in A is adjacent to a vertex in B. Suppose further that if $G \in \mathcal{G}_\alpha(c)$, then every subgraph of G is in $\mathcal{G}_\alpha(c)$. Then any $G \in \mathcal{G}_\alpha(c)$ can be partitioned*

$$V(G) = A' \cup B' \cup C',$$

where $|A'|, |B'| \leq (1/2)|V(G)|$,

$$|C'| \leq \frac{c}{1 - (2/3)^\alpha} |V(G)|^\alpha,$$

and no vertex in A' is adjacent to a vertex in B'.

Proof. We follow the proof of Corollary 3 of [9]. We inductively define a sequence of sets $\{A_i, B_i, C_i, D_i\}$ so that $V(G) = A_i \cup B_i \cup C_i \cup D_i$ is a vertex partition; there are no edges between any of the sets A_i, B_i, and D_i, and we have $|A_i| \leq |B_i| \leq |A_i \cup C_i \cup D_i|$ and $|D_i| \leq \frac{2}{3}|D_{i-1}|$.

Let $A_0 = B_0 = C_0 = \emptyset$ and $D_0 = V(G)$. The properties above are clearly satisfied for these sets. Assume that $A_{i-1}, B_{i-1}, C_{i-1}, D_{i-1}$ have been defined satisfying the above properties and further assume that $D_{i-1} \neq \emptyset$. Applying the hypotheses on $\mathcal{G}_\alpha(c)$ to the graph $G[D_{i-1}]$, we have $D_{i-1} = \widetilde{A} \cup \widetilde{B} \cup \widetilde{C}$, where $|\widetilde{A}|, |\widetilde{B}| \leq \frac{2}{3}|D_{i-1}|$, $|\widetilde{C}| \leq c|D_{i-1}|^\alpha$, and no vertex in $\widetilde{A}$ is adjacent to a vertex in $\widetilde{B}$. We can assume that $|\widetilde{A}| \leq |\widetilde{B}|$.

Let A_i be the smaller (in cardinality) of the sets $A_{i-1} \cup \widetilde{A}$ and B_{i-1}. Let B_i be the other set. Let $C_i = C_{i-1} \cup \widetilde{C}$, and let $D_i = \widetilde{B}$. Then

$$\begin{aligned} A_i \cup B_i \cup C_i \cup D_i &= A_{i-1} \cup \widetilde{A} \cup B_{i-1} \cup C_{i-1} \cup \widetilde{C} \cup \widetilde{B} \\ &= A_{i-1} \cup B_{i-1} \cup C_{i-1} \cup D_{i-1} \\ &= V(G). \end{aligned}$$

Since no vertex in A_{i-1} is adjacent to one in B_{i-1} or $\widetilde{B}$, then no vertex in A_i is adjacent to one in B_i or D_i. Similarly, no vertex in B_i is adjacent to one in D_i.

Also, $|A_i| \leq |B_i|$ by our choice of A_i, and if $A_i = B_{i-1}$, then $B_i = A_{i-1} \cup \widetilde{A}$ and

$$\begin{aligned} |A_i \cup C_i \cup D_i| &\geq \left| B_{i-1} \cup \widetilde{B} \right| \\ &\geq \left| A_{i-1} \cup \widetilde{A} \right| = |B_i|. \end{aligned}$$

If $A_i = A_{i-1} \cup \widetilde{A}$, then $B_i = B_{i-1}$ and

$$\begin{aligned} |A_i \cup C_i \cup D_i| &\geq |A_{i-1} \cup C_{i-1} \cup D_{i-1}| \\ &\geq |B_{i-1}| = |B_i|. \end{aligned}$$

Also, $|D_i| = |\widetilde{B}| \leq \frac{2}{3}|D_{i-1}|$ by the hypotheses. It follows that each term in this sequence of subsets of $V(G)$ satisfies all of the above properties.

As the vertex set of G is finite and $|D_i|$ is decreasing, then $|D_k| = 0$ for some k. Thus for such k, we have $V(G) = A_k \cup B_k \cup C_k$, $|A_k| \leq |B_k| \leq |A_k \cup C_k|$, and no vertex in A_k is adjacent to one in B_k. Let $A' = A_k$, $B' = B_k$, and $C' = C_k$. It follows that $|A'|, |B'| \leq n/2$. Now,

$$\begin{aligned} |C_i| &= |C_{i-1}| + \left|\widetilde{C}\right| \\ &\leq |C_{i-1}| + c|D_{i-1}|^\alpha. \end{aligned}$$

As $|D_0| = n$, then $|D_i| \leq (2/3)^i n$, so

$$|C_i| \leq |C_{i-1}| + c(2/3)^{(i-1)\alpha} n^\alpha.$$

As $|C_i| \leq c(2/3)^{(1-1)\alpha} n^\alpha + c(2/3)^{(2-1)\alpha} n^\alpha + \cdots + c(2/3)^{(i-1)\alpha} n^\alpha$, then

$$\begin{aligned} |C'| &\leq \sum_{i=0}^{\infty} \left(\frac{2}{3}\right)^{i\alpha} cn^\alpha \\ &= \frac{c}{1-(2/3)^\alpha}\, n^\alpha. \qquad \square \end{aligned}$$

THEOREM 2.2 (Alon, Seymour, Thomas (1990) [1]). *Let G be a graph with n vertices and no K_h-minor, for fixed $h \geq 3$. Then there exists a partition $V(G) = A \cup B \cup C$ such that $|A|, |B| \leq 2n/3$, $|C| \leq h^{3/2} n^{1/2}$, and no vertex in A is adjacent to a vertex in B.*

A straightforward application of Proposition 2.1 and Theorem 2.2 gives the following.

COROLLARY 2.3. *Let G be a graph with n vertices and no K_h-minor, for fixed $h \geq 3$. Then $V(G) = A \cup B \cup C$, where $|A|, |B| \leq n/2$, $|C| \leq \frac{h^{3/2}}{1-\sqrt{2/3}}\, n^{1/2}$, and no vertex in A is adjacent to a vertex in B.*

The well-known separation theorem for planar graphs [9] was improved in [2] to give the best known such result thus far. See also [6] for results on the decomposition of planar graphs.

THEOREM 2.4 (Alon, Seymour, Thomas (1994) [2]). *Let G be a planar graph with n vertices. Then there exists a partition $V(G) = A \cup B \cup C$ such that $|A|, |B| \leq 2n/3$, $|C| \leq 3\sqrt{2}/2\; n^{1/2}$, and no vertex in A is adjacent to a vertex in B.*

A straightforward application of Proposition 2.1 and Theorem 2.4 gives the following.

COROLLARY 2.5. *Let G be a planar graph with n vertices. Then there exists a partition $V(G) = A \cup B \cup C$ such that $|A|, |B| \leq n/2$, $|C| \leq \frac{3\sqrt{2}}{2(1-\sqrt{2/3})}\, n^{1/2}$, and no vertex in A is adjacent to a vertex in B.*

The separation theorem for planar graphs in [9] was generalized in [7, 5] to graphs with a fixed genus g. Below is the separator theorem from [5], which is slightly stronger than the theorem in [7].

THEOREM 2.6 (Djidjev (1985) [5]). *Let G be a graph with n vertices and genus g. Then there exists a partition $V(G) = A \cup B \cup C$ such that $|A|, |B| \leq 2n/3$, $|C| \leq \sqrt{6(2g+1)n}$, and no vertex in A is adjacent to a vertex in B.*

A straightforward application of Proposition 2.1 and Theorem 2.6, together with the observation that a subgraph of a graph of genus g has a genus $\leq g$, gives the following.

COROLLARY 2.7. *Let G be a graph with n vertices and genus g. Then there exists a partition $V(G) = A \cup B \cup C$ such that $|A|, |B| \leq n/2$, $|C| \leq \frac{\sqrt{6(2g+1)}}{1-\sqrt{2/3}}\, n^{1/2}$, and no vertex in A is adjacent to a vertex in B.*

3. Upper bounds on the integrity of graphs. In this section we give upper bounds for the integrity of certain graphs, which have a separator theorem of the type given in Corollaries 2.3, 2.5, and 2.7.

THEOREM 3.1. *Let $0 \le \alpha < 1$ and $1 \le c$. Let $\mathcal{G}_\alpha(c)$ be a family of graphs so that for any $G \in \mathcal{G}_\alpha(c)$, with $|V(G)| = n$, there exists a partition $V(G) = A \cup B \cup C$ such that $|A|, |B| \le n/2$, $|C| \le cn^\alpha$, and no vertex in A is adjacent to a vertex in B. Further, assume that if $G \in \mathcal{G}_\alpha(c)$, then every subgraph of G is in $\mathcal{G}_\alpha(c)$. Then for any $G \in \mathcal{G}_\alpha(c)$ and $n \ge (2c)^{1/(1-\alpha)}$, we have*

$$I(G) \le an^{\frac{1}{2-\alpha}} - bn^\alpha,$$

where

$$a = c^{\frac{1}{2-\alpha}} 2^{-\frac{1-\alpha}{2-\alpha}} \left(1 + \frac{1}{1 - 2^{-(1-\alpha)}}\right)$$

and

$$b = \frac{c}{2^{1-\alpha} - 1}.$$

Note that $\frac{1}{2-\alpha} > \alpha$ for $0 \le \alpha < 1$.

Proof. Let $G \in \mathcal{G}_\alpha(c)$, with $|V(G)| = n$. Then $V(G) = A \cup B \cup C$ by the hypothesis. By removing the set of vertices C, we divide G into components $G[A]$ and $G[B]$, each of which has no more than $n/2$ vertices. This directly gives the estimate

$$I(G) \le \frac{n}{2} + cn^\alpha.$$

Now we apply the separator theorem to each of the subgraphs $G[A]$ and $G[B]$. Thus

$$A = A_1 \cup B_1 \cup C_1,$$

where $|A_1|, |B_1| \le n/4$, and $|C_1| \le c(n/2)^\alpha$ and similarly for $B = A_2 \cup B_2 \cup C_2$. By removing the vertices in C_1 and C_2, we decompose G into 4 components, each with no more than $n/4$ vertices. It follows that

$$\begin{aligned} I(G) &\le \frac{n}{4} + |C| + |C_1| + |C_2| \\ &\le \frac{n}{4} + cn^\alpha + 2c\left(\frac{n}{2}\right)^\alpha. \end{aligned}$$

Continuing in this way, we apply the separator theorem successively ℓ times (where ℓ is a nonnegative integer to be specified later). At each step, we remove vertices to separate each of 2^i components already obtained with a set of vertices of size no more than $c(n/2^i)^\alpha$. After i steps, we have decomposed G into 2^i components, each containing no more than $n/2^i$ vertices. At the ith-step we would remove no more than $2^{i-1}c(n/2^{i-1})^\alpha$ vertices.

It follows that for any nonnegative integer ℓ, we have the estimate

$$I(G) \le \frac{n}{2^\ell} + \sum_{i=0}^{\ell-1} c2^i \Big(\frac{n}{2^i}\Big)^\alpha \tag{3.1}$$

$$= n + \sum_{i=0}^{\ell-1} \Big(c2^i \Big(\frac{n}{2^i}\Big)^\alpha - \frac{n}{2^{i+1}}\Big). \tag{3.2}$$

Now we set a value for ℓ. Define

$$\ell = \max\left(0, 1 + \left\lceil \log_2 \left(\frac{n^{1-\alpha}}{2c}\right)^{\frac{1}{2-\alpha}} \right\rceil\right).$$

(This value of ℓ minimizes (3.2). This follows from the fact that

$$c2^i \left(\frac{n}{2^i}\right)^{\alpha} - \frac{n}{2^{i+1}} \leq 0$$

if and only if $i \leq \log_2(\frac{n^{1-\alpha}}{2c})^{1/(2-\alpha)}$.)

Since $n \geq (2c)^{1/(1-\alpha)}$ implies $1 \leq n^{1-\alpha}/(2c)$, we get

$$\ell = 1 + \left\lceil \log_2 \left(\frac{n^{1-\alpha}}{2c}\right)^{\frac{1}{2-\alpha}} \right\rceil \tag{3.3}$$

$$= 1 + \log_2 \left(\frac{n^{1-\alpha}}{2c}\right)^{\frac{1}{2-\alpha}} - \log_2(1/\delta), \tag{3.4}$$

where $\log_2(1/\delta) \in [0, 1)$ is the fractional part of the second term in (3.3). Thus

$$2^{\ell} = 2\left(\frac{n^{1-\alpha}}{2c}\right)^{\frac{1}{2-\alpha}} \delta \quad \text{for some} \quad \delta \in \left(\frac{1}{2}, 1\right].$$

Substituting this expression into the right-hand side of the estimate

$$I(G) \leq \frac{n}{2^{\ell}} + \sum_{i=0}^{\ell-1} c2^i \left(\frac{n}{2^i}\right)^{\alpha} = \frac{n}{2^{\ell}} + cn^{\alpha}\frac{(2^{1-\alpha})^{\ell} - 1}{2^{1-\alpha} - 1},$$

we get

$$I(G) \leq (cn)^{\frac{1}{2-\alpha}} 2^{-\frac{1-\alpha}{2-\alpha}} \delta^{-1} + \frac{(cn)^{\frac{1}{2-\alpha}} 2^{\frac{(1-\alpha)^2}{2-\alpha}}}{2^{1-\alpha} - 1}\delta^{1-\alpha} - \frac{c}{2^{1-\alpha} - 1} n^{\alpha}.$$

Let $f(\delta)$ denote the right-hand side of the above inequality. Straightforward calculations show that $\lim_{\delta \to 0^+} f(\delta) = \lim_{\delta \to +\infty} f(\delta) = +\infty$, and the only critical point of $f(\delta)$ on $(0, +\infty)$ is

$$\delta = \left(\frac{1 - 2^{-(1-\alpha)}}{1 - \alpha}\right)^{\frac{1}{2-\alpha}}.$$

Furthermore, this critical point is in $(1/2, 1]$ for any $\alpha \in [0, 1)$. It follows that

$$\sup_{\delta \in (\frac{1}{2}, 1]} f(\delta) = \max\left(f\left(\frac{1}{2}\right), f(1)\right).$$

It is easy to verify that, in fact, $f(1/2) = f(1)$. Hence

$$I(G) \leq f(1) = (cn)^{\frac{1}{2-\alpha}} 2^{-\frac{1-\alpha}{2-\alpha}} \left(1 + \frac{1}{1 - 2^{-(1-\alpha)}}\right) - \frac{c}{2^{1-\alpha} - 1} n^{\alpha}. \qquad \square$$

Example 3.2. Let G be a graph which is the union of finitely many paths, and let $n = |V(G)|$. Theorem 3.1 now implies (with $\alpha = 0$ and $c = 1$) that $I(G) \leq \frac{3}{2}\sqrt{2n} - 1$. Note that this bound is quite sharp, as for the path P_n of length n, we have $I(P_n) = \lceil 2\sqrt{n+1} \rceil - 2$ and $\frac{3}{2}\sqrt{2} \approx 2.121$. As paths have the maximum integrity of trees of order n, see Lemma 5 from [10], we get $I(G) \leq \frac{3}{2}\sqrt{2n} - 1$ for G a tree.

Proof of Theorem 1.1. Let G be a graph with n vertices and no K_h-minor. Set $c = \frac{h^{3/2}}{1-\sqrt{2/3}}$ and $\alpha = 0.5$. Using Corollary 2.3 and Theorem 3.1, we get that for $n \geq 119h^3$, we have

$$I(G) \leq 10.9hn^{2/3} - 13.1h^{3/2}n^{1/2}. \qquad \square$$

Proof of Theorem 1.2. Let G be a planar graph with n vertices. Set $c = \frac{3\sqrt{2}}{2(1-\sqrt{2/3})}$ and $\alpha = 0.5$. Using Corollary 2.5 and Theorem 3.1 we get that for $n \geq 535$, we have

$$I(G) \leq 18n^{2/3} - 27.9n^{1/2}. \qquad \square$$

Proof of Theorem 1.3. Let G be a graph with n vertices and genus at most g. Define $c = \frac{\sqrt{6(2g+1)}}{1-\sqrt{2/3}}$ and $\alpha = 0.5$. Using Corollary 2.7 and Theorem 3.1, we gain that for $n \geq 713(2g+1)$, we have

$$I(G) \leq 19.8(2g+1)^{1/3}n^{2/3} - 32.2(2g+1)^{1/2}n^{1/2}. \qquad \square$$

4. Rectangular boxes in the lattice graph $\mathbb{Z}^d$. Let d be a positive integer. Recall that a subgraph G of $\mathbb{Z}^d$, which forms a rectangular box that is parallel to the axes, is called a box-graph. We say that G has dimensions $a_1, \ldots, a_d$ if $a_1 \geq \cdots \geq a_d$, $a_i \in \mathbb{N}$, and G contains all vertices $(x_1, \ldots, x_d)$, where $x_i \in \mathbb{Z}$ and $0 \leq x_i < a_i$ for all i. Let m_d denote the number of vertices on a smallest hyperface of a box-graph G. Then $m_d = \Pi_{i=2}^{d} a_i$. In the case of $d = 1$, we define $m_1 = 1$.

The following lemma asserts the (seemingly obvious) statement that if we delete a set of vertices S from a box-graph G and $|S|$ is small, then $G \setminus S$ will contain a large component.

LEMMA 4.1. *Let G be a box-graph in $\mathbb{Z}^d$. For any $\epsilon \in (0,1)$, there exists $c_d \in (0,1)$ such that if $S \subset G$ and $|S| \leq c_d m_d$, then there exists a component K of $G \setminus S$ such that $|K| \geq (1-\epsilon)|V(G)|$.*

Proof. The proof is by induction on the dimension d. We observe that the lemma is true for $d = 1$, since $|S| \leq c_1 m_1$, with $c_1 \in (0,1)$ implies $S = \emptyset$.

Let $d \geq 2$ and assume that our statement is true for $d-1$. Let c_d be a small number (which we specify later). Let a_1 be the largest dimension of the box-graph G, and let e be an "edge" of the box consisting of a_1 vertices lying on a line. (Here the word "edge" is used as in geometry, not in graph theory.) Set $u = a_1$.

Consider the u cross sections of G which are orthogonal to e. Observe that, by the choice of e, each such cross section consists of m_d vertices. Denote by $\mathcal{H}$ the set of those cross sections which have at most $\sqrt{c_d} m_d/u$ vertices from S. Note that $|S| \leq c_d m_d$ implies that there are less than $\sqrt{c_d}u$ cross sections not in $\mathcal{H}$, and therefore $|\mathcal{H}| > (1 - \sqrt{c_d})u$. Let $R \in \mathcal{H}$. We regard R as a $(d-1)$ dimensional box-graph with $|R| = m_d$ vertices. Let m_{d-1} denote the minimal number of vertices on any of the $(d-2)$ dimensional faces of R.

Let $C \in (0,1)$, which we specify later. Note that, by the definition of u, we have $m_d/u \leq m_{d-1}$ (which is true even for $d = 2$), thus $|S \cap R| \leq \sqrt{c_d} m_{d-1}$. By

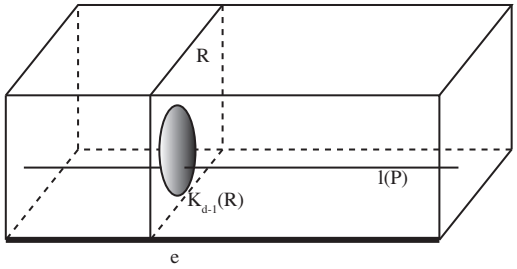


FIG. 4.1. *A schematic representation of a box-graph, with a cross section R and the component $K_{d-1}(R)$. Lines of the type $l(P)$ connect many such components from different cross sections.*

the induction hypotheses, if c_d is small enough, then *for any* $R \in \mathcal{H}$, there exists a component $K_{d-1}(R)$ of $R \setminus S$ such that

$$|K_{d-1}(R)| \geq Cm_d. \tag{4.1}$$

For an illustration, see Figure 4.1. Our goal is to construct the desired component K of $G \setminus S$ as a union of certain components $K_{d-1}(R)$ to yield the desired size.

For a vertex $P \in V(G)$, let $l(P)$ denote those vertices of G which are on the line passing through P and parallel to e. Let

$$G' := \{P \in V(G) \mid l(P) \cap S = \emptyset\}$$

and $T' := T \cap G'$, where T is an arbitrary fixed cross section of G orthogonal to e. Let

$$J := \cup_{R \in \mathcal{H}} (K_{d-1}(R) \cap G').$$

Since

$$|K_{d-1}(R) \cap G'| \geq |K_{d-1}(R)| - |S| \geq (C - c_d)m_d,$$

we have $|J| > (1 - \sqrt{c_d})u(C - c_d)m_d$. J may not be a connected subset of $G \setminus S$. To obtain a connected subset of J, we proceed as follows: $|T'| \leq m_d$ implies that there exists $P \in T'$ such that $|l(P) \cap J| > (1 - \sqrt{c_d})u(C - c_d)$. Thus the line segment $l(P)$ connects more than $(1 - \sqrt{c_d})u(C - c_d)$ of the sets $K_{d-1}(R)$ $(R \in \mathcal{H})$. Hence, there exists a component K of $G \setminus S$ such that

$$|K| > (1 - \sqrt{c_d})u(C - c_d)Cm_d = (1 - \sqrt{c_d})(C - c_d)C|V(G)|.$$

To complete the proof, we choose $C \in (0, 1)$ to be so close to 1 and then c_d to be so close to 0 such that (4.1) holds for any $R \in \mathcal{H}$ and

$$(1 - \sqrt{c_d})(C - c_d)C \geq 1 - \epsilon. \qquad \square$$

Proof of Theorem 1.4. We show that Theorem 1.4 follows from the following lemma.

LEMMA 4.2. *Let G be a box-graph in $\mathbb{Z}^d$, with dimensions $a_1, \ldots, a_d$, where $a_1 \geq \cdots \geq a_d$. Let $A_0 = 1$ and $A_m = (a_1 \ldots a_m)^{1/(m+1)}$ for $m \in \{1, \ldots, d\}$, and let*

$$\mathcal{N} := \Big\{ m \in \{0, \ldots, d-1\} \;\Big|\; a_{m+1} < 2A_m \Big\}.$$

If $\mathcal{N}$ is nonempty, then let $N := \min \mathcal{N}$, otherwise let $N := d$. Then

$$c_d^* \frac{|V(G)|}{A_N} \le I(G) \le C_d^* \frac{|V(G)|}{A_N}, \tag{4.2}$$

where the constants c_d^* and C_d^* depend on d only.

Intuitively, we can explain (4.2) as follows. The sides a_m, $m \ge N+1$ are too small relative to the bigger sides, and this means that the box is flat in dimensions $N+1, \dots, d$ and basically, it has N "real dimensions." In the formula (4.2), $a_{N+1}, \dots, a_d$ will be on the first power (in $|V(G)|/A_N$), whereas the powers of the first N "real dimensions" $a_1, \dots, a_N$ will be less than one. The first N dimensions $a_1, \dots, a_N$ have to be cut by hyperplanes to achieve the integrity bound.

We now show that (4.2) is equivalent to (1.1). By the definition of N, we have $a_{m+1} \ge 2A_m$ for $m = 0, 1, \dots, N-1$, and if $N < d$, we also have $a_{N+1} < 2A_N$.

Note that $a_{m+1} \ge 2A_m = 2(a_1 \dots a_m)^{1/(m+1)}$ implies $a_1 \dots a_{m+1} \ge 2(a_1 \dots a_m)^{1+1/(m+1)}$, and so we have $(a_1 \dots a_{m+1})^{1/(m+2)} \ge b_m (a_1 \cdots a_m)^{1/(m+1)}$, where $b_m = 2^{1/(m+2)} > 1$. Thus $A_{m+1} \ge A_m$ holds for $m = 0, 1, \dots, N-1$.

Also, $a_{N+1} < 2A_N = 2(a_1 \dots a_N)^{1/(N+1)}$ implies $a_{N+1}^{1+1/(N+1)} < 2(a_1 \dots a_{N+1})^{1/(N+1)}$, so $a_{N+2} \le a_{N+1} < 2(a_1 \dots a_{N+1})^{1/(N+2)}$. Continuing in this way, we get that $a_{m+1} < 2(a_1 \dots a_m)^{1/(m+1)}$ holds for $m = N, \dots, d-1$. As in the previous paragraph, this implies that $A_{m+1} < b_m A_m$ holds for $m = N, \dots, d-1$.

We conclude that $b_{d-1} b_{d-2} \dots b_N A_N \ge A_m$ for all $m = 0, \dots, d$. Here

$$b_{d-1} b_{d-2} \dots b_N \le 2^{1/2+1/3+\cdots+1/(d+1)} \le 2d,$$

and so

$$\frac{|V(G)|}{2d^2 A_N} \le \frac{|V(G)|}{A_1 + \cdots + A_d} \le \frac{|V(G)|}{A_N},$$

establishing the equivalence of (4.2) and (1.1). ∎

Proof of Lemma 4.2. First we prove the lower bound in (4.2). We can assume that $1 \le N$ (otherwise $a_1 < 2A_0$, and so $1 = a_1 = \cdots = a_d$ and G is a single vertex). Then we have $2 \le a_1$. By definition, when $\mathcal{N} \ne \emptyset$, we have $a_{N+1} < 2A_N$ and $a_N \ge 2A_{N-1}$, whereas when $\mathcal{N} = \emptyset$, we have $a_N \ge 2A_{N-1}$ (and $N = d$). Note that this implies

$$\begin{aligned} a_N &= a_N \frac{(a_1 \dots a_N)^{\frac{1}{N+1}}}{(a_1 \dots a_N)^{\frac{1}{N+1}}} = \frac{a_N^{1-\frac{1}{N+1}}}{(a_1 \dots a_{N-1})^{\frac{1}{N+1}}} A_N \\ &= \frac{a_N^{\frac{N}{N+1}}}{\left((a_1 \dots a_{N-1})^{\frac{1}{N}}\right)^{\frac{N}{N+1}}} A_N = \left(\frac{a_N}{A_{N-1}}\right)^{\frac{N}{N+1}} A_N \\ &\ge 2^{\frac{N}{N+1}} A_N \ge \sqrt{2} A_N. \end{aligned}$$

So $[\frac{a_i}{A_N}]$ $(i = 1, \dots, N)$ are positive integers.

Let $k_i \in \{0, \dots, [\frac{a_i}{A_N}]\}$ $(i = 1, \dots, N)$ and consider the box-graph

$$\begin{aligned} B(k_1, \dots, k_N) := &\left\{x_1 \in \mathbb{Z} : \ [k_1 A_N] < x_1 \le \min([(k_1+1)A_N], a_1)\right\} \\ &\times \cdots \times \left\{x_N \in \mathbb{Z} : \ [k_N A_N] < x_N \le \min([(k_N+1)A_N], a_N)\right\} \\ \times &\left\{x_{N+1} \in \mathbb{Z} : \ 0 < x_{N+1} \le a_{N+1}\right\} \times \cdots \times \left\{x_d \in \mathbb{Z} : \ 0 < x_d \le a_d\right\}, \end{aligned}$$

which, for simplicity, we call a box. When $k_i = [\frac{a_i}{A_N}]$ holds for at least one $i \in \{0, \dots, N\}$, we call $B(k_1, \dots, k_N)$ a truncated box (it may even be an empty set), and otherwise we call it a full box. Clearly, the $B(k_1, \dots, k_N)$ are disjoint sets, and their union is $V(G)$. Since $A_N - 1 < [(k_i + 1)A_N] - [k_i A_N] < A_N + 1$, any of the first N dimensions of a full box is an integer in the interval $(A_N - 1, A_N + 1)$. The remaining dimensions of a full box are $a_{N+1}, \dots, a_d$.

The number of full boxes is $\prod_{i=1}^{N} [\frac{a_i}{A_N}]$. Since $A_N \le a_i/\sqrt{2}$ $(i = 1, \dots, N)$, we have

$$\begin{aligned} A_N &= \prod_{i=1}^{N} \frac{a_i}{A_N} \ge \prod_{i=1}^{N} \left[\frac{a_i}{A_N} \right] \\ &\ge \prod_{i=1}^{N} \left(\frac{a_i}{A_N} - 1 \right) \ge \frac{\prod_{i=1}^{N} \left(a_i - \frac{1}{\sqrt{2}} a_i \right)}{A_N^N} \ge \left(1 - \frac{1}{\sqrt{2}} \right)^d A_N. \end{aligned} \tag{4.3}$$

Let M denote the maximum dimension (i.e., maximal number of vertices on an edge parallel to a coordinate axis) of any of the full boxes. When $\mathcal{N} \neq \emptyset$, using the estimate $2A_N > A_N + 1$ for the first N dimensions and $2A_N > a_{N+1} \ge \cdots \ge a_d$ for the rest of the dimensions, we get $2A_N > M$. When $\mathcal{N} = \emptyset$, using $2A_N > A_N + 1$ (and $N = d$), we get again $2A_N > M$. The minimal number m_d of vertices on any hyperface of an arbitrary full box is at least

$$\begin{aligned} m_d \ge \frac{(A_N - 1)^N a_{N+1} \dots a_d}{M} &> \frac{(A_N - 1)^N a_{N+1} \dots a_d}{2A_N} \\ &\ge \frac{1}{2} \left(1 - \frac{1}{\sqrt[d+1]{2}} \right)^d A_N^{N-1} a_{N+1} \dots a_d, \end{aligned}$$

where we used that the number of vertices in a full box is at least

$$(A_N - 1)^N a_{N+1} \dots a_d \tag{4.4}$$

and

$$1 - \frac{1}{A_N} \ge 1 - \frac{1}{\sqrt[d+1]{2}}.$$

The last inequality follows from $A_N \ge \sqrt[d+1]{2}$.

Now let $S \subset V(G)$ be arbitrary. Let $0 < \epsilon < 1$ be arbitrary, and let c_d be the number given in the statement of Lemma 4.1.

Case 1. If there exists a full box B such that

$$|S \cap B| \le c_d \frac{1}{2} \left(1 - \frac{1}{\sqrt[d+1]{2}} \right)^d A_N^{N-1} a_{N+1} \dots a_d \le c_d m_d, \tag{4.5}$$

then, by Lemma 4.1 (and (4.4)), we have

$$I(G) \ge (1 - \epsilon)(A_N - 1)^N a_{N+1} \dots a_d \ge (1 - \epsilon) \left(1 - \frac{1}{\sqrt[d+1]{2}} \right)^d A_N^N a_{N+1} \dots a_d. \tag{4.6}$$

Case 2. If

$$|S \cap B| > c_d \frac{1}{2} \left(1 - \frac{1}{\sqrt[d+1]{2}} \right)^d A_N^{N-1} a_{N+1} \dots a_d \tag{4.7}$$

for any full box B, then, by (4.3),

$$I(G) \geq |S| > (\text{number of full boxes}) \cdot c_d \frac{1}{2} \left(1 - \frac{1}{\sqrt[d+1]{2}}\right)^d A_N^{N-1} a_{N+1} \dots a_d$$

$$\geq \left(1 - \frac{1}{\sqrt{2}}\right)^d c_d \frac{1}{2} \left(1 - \frac{1}{\sqrt[d+1]{2}}\right)^d A_N^N a_{N+1} \dots a_d. \tag{4.8}$$

Since $A_N^N a_{N+1} \dots a_d = (\prod_{i=1}^d a_i)/A_N = |V(G)|/A_N$, (4.6) and (4.8) establish the lower bound at (4.2).

To prove the upper bound at (4.2), intersect G with hyperplanes to define the boxes in the first half of the proof. More precisely, let

$$\begin{aligned}
H_1 &:= \left\{ [k_1 A_N] : \ k_1 = 1, \dots, \left[\frac{a_1}{A_N}\right] \right\} \times \{1, \dots, a_2\} \times \cdots \times \{1, \dots, a_d\}, \\
H_2 &:= \{1, \dots, a_1\} \times \left\{ [k_2 A_N] : \ k_2 = 1, \dots, \left[\frac{a_2}{A_N}\right] \right\} \\
&\quad \times \{1, \dots, a_3\} \times \cdots \times \{1, \dots, a_d\}, \\
&\ \ \vdots \\
H_N &:= \{1, \dots, a_1\} \times \cdots \times \{1, \dots, a_{d-1}\} \times \left\{ [k_N A_N] : \ k_N = 1, \dots, \left[\frac{a_N}{A_N}\right] \right\},
\end{aligned}$$

and let $S := \cup_{i=1}^N H_i$. Now,

$$|S| \leq \sum_{i=1}^N |H_i| = \sum_{i=1}^N \left[\frac{a_i}{A_N}\right] \frac{|V(G)|}{a_i} \leq d \frac{|V(G)|}{A_N}.$$

We have seen that any of the first N dimensions of a full box is an integer less than $A_N + 1$, and the next dimensions are $a_{N+1}, \dots, a_d$. Note that this is also true for the truncated boxes (which easily follows from $[([\frac{a_i}{A_N}]+1)A_N] \geq a_i, \ i = 1, \dots, N$). So any box (and hence any component of $G \setminus S$) has at most $(A_N + 1)^N a_{N+1} \dots a_d$ vertices. These lead to

$$\begin{aligned}
I(G) &\leq d \frac{|V(G)|}{A_N} + (A_N + 1)^N a_{N+1} \dots a_d \\
&\leq d \frac{|V(G)|}{A_N} + \left(1 + \frac{1}{\sqrt[d+1]{2}}\right)^d A_N^N a_{N+1} \dots a_d \leq \left(d + \left(1 + \frac{1}{\sqrt[d+1]{2}}\right)^d\right) \frac{|V(G)|}{A_N},
\end{aligned} \tag{4.9}$$

where we also used

$$1 + \frac{1}{A_N} \leq 1 + \frac{1}{\sqrt[d+1]{2}}. \qquad \square$$

It is possible to give values to the constants c_d^* and C_d^* in Lemma 4.2. For example, below we consider the case $d = 2$. (Note that one can certainly find better constants for the upper and lower bounds if one considers a proof that applies to the special case $d = 2$ directly.)

Proof of Theorem 1.5. Let $d = 2$. First we give concrete values for the constants in Lemma 4.1. Let $\epsilon \in (0, 1)$. Since in the proof of Lemma 4.1 we have that $c_1 \in (0, 1)$

may be chosen arbitrarily and that C can be as close to one as we wish, $c_2 \in (0,1)$ can be any number satisfying

$$(1-\sqrt{c_2})(1-c_2) > 1-\epsilon.$$

Furthermore, we want to choose ϵ and c_2 to get about the same lower bounds in (4.6) and (4.8). Thus we desire

$$1-\epsilon \approx \frac{1}{2}c_2\left(1-\frac{1}{\sqrt{2}}\right)^2.$$

Set $\epsilon = 0.968$ and $c_2 = 0.754$. Now, (4.6) and (4.8) give $c_2^* = 0.00136$. Together with the upper bound from (4.9), we have

$$0.00136\frac{|V(G)|}{A_N} \le I(G) \le 5.22\frac{|V(G)|}{A_N}. \tag{4.10}$$

(This holds even in the case when $N = 0$.)

Now in Lemma 4.2, we have $A_0 = 1$, $A_1 = \sqrt{a_1}$, and $A_2 = \sqrt[3]{a_1 a_2}$. Note that if $a_1 < 2A_0 = 2$, then necessarily $a_1 = a_2 = 1$, and the claim of Theorem 1.5 is satisfied. So we may assume that $a_1 \ge 2A_0$.

If $a_2 \ge 2A_1 = 2\sqrt{a_1}$, then $N = 2$ in Lemma 4.2, and so (4.2) leads to

$$0.00136n^{2/3} \le I(G) \le 5.22n^{2/3}.$$

If $a_2 < 2A_1 = 2\sqrt{a_1}$, then $N = 1$ in Lemma 4.2, and so (4.2) leads to

$$0.00136\sqrt{a_1}a_2 \le I(G) \le 5.22\sqrt{a_1}a_2. \qquad \square$$

Note that the above inequality gives $I(G) = O(\sqrt{n})$ in the degenerate cases when $a_2 = O(1)$. This is consistent with the known integrity of a path ($a_2 = 1$); see [3].

Proof of Theorem 1.6. Let $G \subset \mathbb{Z}^d$ be a cube graph whose dimensions are of size a. Note that the inequality $a_{m+1} < 2A_m$ in Lemma 4.2 now simplifies to $a < 2^{m+1}$. Thus, for $a \ge 2^{m+1}$, we have $\mathcal{N} = \emptyset$, and so $N = d$. Since $|V(G)|/A_d = a^d/a^{d/(d+1)} = a^{d^2/(d+1)} = |V(G)|^{d/(d+1)}$, Lemma 4.2 gives

$$c_d^*|V(G)|^{d/(d+1)} \le I(G) \le C_d^*|V(G)|^{d/(d+1)},$$

which holds even in the case $a < 2^{m+1}$ if we redefine the values of the constants c_d^* and C_d^*. $\square$

REFERENCES

[1] N. Alon, P. Seymour, and R. Thomas, *A separator theorem for nonplanar graphs,* J. Amer. Math. Soc., 3 (1990), pp. 801–808.

[2] N. Alon, P. Seymour, and R. Thomas, *Planar separators,* SIAM J. Discrete Math., 7 (1994), pp. 184–193.

[3] K.S. Bagga, L.W. Beineke, W.D. Goddard, M.J. Lipman, and R.E. Pippert, *A survey of integrity,* Discrete Appl. Math., 37/38 (1992), pp. 13–28.

[4] C.A. Barefoot, R. Entringer, and H. Swart, *Vulnerability in graphs—A comparative survey,* J. Combin. Math. Combin. Comput., 1 (1987), pp. 12–22.

[5] H.N. Djidjev, *A separator theorem for graphs of fixed genus,* Serdica Math. J., 11 (1985), pp. 319–329.

[6] F.V. Fomin and D.M. Thilikos, *New upper bounds on the decomposability of planar graphs,* J. Graph Theory, 51 (2006), pp. 53–81.

[7] 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.
[8] W. Goddard and H. Swart, *Integrity in graphs: bounds and basics,* J. Combin. Math. Combin. Comput., 7 (1990), pp. 139–151.
[9] R.J. Lipton and R.E. Tarjan, *A separator theorem for planar graphs,* SIAM J. Appl. Math., 36 (1979), pp. 177–189.
[10] A. Vince, *The integrity of a cubic graph,* Discrete Appl. Math., 140 (2004), pp. 223–239.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 278–287

CAN A GRAPH HAVE DISTINCT REGULAR PARTITIONS?*

NOGA ALON†, ASAF SHAPIRA‡, AND URI STAV§

Abstract. The regularity lemma of Szemerédi gives a concise approximate description of a graph via a so-called regular partition of its vertex set. In this paper we address the following problem: Can a graph have two "distinct" regular partitions? It turns out that (as observed by several researchers) for the standard notion of a regular partition, one can construct a graph that has very distinct regular partitions. On the other hand, we show that for the stronger notion of a regular partition that has been recently studied, all such regular partitions of the same graph must be very "similar." En route, we also give a short argument for deriving a recent variant of the regularity lemma obtained independently by Rödl and Schacht and by Lovász and Szegedy from a previously known variant of the regularity lemma due to Alon et al. in 2000. The proof also provides a deterministic polynomial time algorithm for finding such partitions.

Key words. regularity lemma, algorithm, isomorphic

AMS subject classifications. 68R01, 05D99

DOI. 10.1137/070695952

1. Introduction.

We start with some of the basic definitions of regularity and state the regularity lemmas that we refer to in this paper. For a comprehensive survey on the regularity lemma, the reader is referred to [7]. For a set of vertices $A \subseteq V$, we denote by $E(A)$ the set of edges of the graph induced by A in G, and by $e(A)$ the size of $E(A)$. Similarly, if $A \subseteq V$ and $B \subseteq V$ are two vertex sets, then $E(A, B)$ stands for the set of edges of G connecting vertices in A and B, and $e(A, B)$ denotes the number of ordered pairs (a, b) such that $a \in A$, $b \in B$, and ab is an edge of G. Note that if A and B are disjoint, then this is simply the number of edges of G that connect a vertex of A with a vertex of B, that is, $e(A, B) = |E(A, B)|$. The *edge density* of the pair (A, B) is defined as $d(A, B) = e(A, B)/|A||B|$. When several graphs on the same set of vertices are involved, we write $d_G(A, B)$ to specify the graph to which we refer.

DEFINITION 1.1 (ϵ-regular pair). *A pair (A, B) is* ϵ-regular, *if for any two subsets $A' \subseteq A$ and $B' \subseteq B$, satisfying $|A'| \geq \epsilon|A|$ and $|B'| \geq \epsilon|B|$, the inequality $|d(A', B') - d(A, B)| \leq \epsilon$ holds.*

A partition $\mathcal{A} = \{V_i : 1 \leq i \leq k\}$ of the vertex set of a graph is called an *equipartition* if $|V_i|$ and $|V_j|$ differ by no more than 1 for all $1 \leq i < j \leq k$ (so, in particular, each V_i has one of two possible sizes). For the sake of brevity, we will henceforth use the term *partition* to denote an equipartition. We call the number of sets in a partition (k above) the *order* of the partition.

*Received by the editors June 30, 2007; accepted for publication (in revised form) July 10, 2008; published electronically January 7, 2009. A preliminary version of this paper appeared in the Proceedings of the 13th International Computing and Combinatorics Conference (COCOON 2007), pp. 428–438.

http://www.siam.org/journals/sidma/23-1/69595.html

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

‡Microsoft Research, One Microsoft Way, Redmond, WA 98052 (asafico@tau.ac.il).

§School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel (uristav@tau.ac.il).

DEFINITION 1.2 (ϵ-regular partition). *A partition $\mathcal{V} = \{V_i : 1 \le i \le k\}$ of $V(G)$ for which all but at most $\epsilon\binom{k}{2}$ of the pairs (V_i, V_j) are ϵ-regular is called an* ϵ-regular partition *of $V(G)$.*

The regularity lemma of Szemerédi can be formulated as follows.

LEMMA 1.3 (see [14]). *For every m and $\epsilon > 0$ there exists an integer $T = T_{1.3}(m, \epsilon)$ with the following property: Any graph G on $n \ge T$ vertices has an ϵ-regular partition $\mathcal{V} = \{V_i : 1 \le i \le k\}$ with $m \le k \le T$.*

The main drawback of Szemerédi's regularity lemma is that the bounds on the integer T, and hence on the order of $\mathcal{V}$, have an enormous dependency on $1/\epsilon$. The current bounds are towers of exponents of height $O(1/\epsilon^5)$. This means that the regularity measure (ϵ in Lemma 1.3) is very large compared to the inverse of the order of the partition (k in Lemma 1.3). In some cases, however, we would like the regularity measure between the pairs to have some (strong) relation to the order of the partition. This leads to the following definition.

DEFINITION 1.4 (f-regular partition). *For a function $f : \mathbb{N} \to (0, 1)$, a partition $\mathcal{V} = \{V_i : 1 \le i \le k\}$ of $V(G)$ is said to be* f-regular *if all pairs (V_i, V_j), $1 \le i < j \le k$, are $f(k)$-regular.*

Note that, as opposed to Definition 1.2, in the above definition the order of the partition and the regularity measure between the sets of the partition go "hand in hand" via the function f. One can (more or less) rephrase Lemma 1.3 as saying that every graph has a $(\log^*(k))^{-1/5}$-regular partition.[1] Furthermore, Gowers [5] showed that this is close to being tight. Therefore, one cannot guarantee that a general graph has an f-regular partition for a function f approaching zero faster than roughly $1/\log^*(k)$. One should thus look for certain variants of this notion and still be able to show that any graph has a similar partition.

A step in this direction was first taken by Alon et al. [2] who proved a stronger variant of the regularity lemma. See Lemma 2.3 for the precise statement. The following is yet another variant of the regularity lemma that was recently proved independently by Rödl and Schacht [11] (where it is called "the regular approximation lemma") and by Lovász [9] (implicitly following a result of Lovász and Szegedy in [10]). This lemma does not guarantee that for any f we can find an f-regular partition of any given graph. Rather, it shows that any graph is "close" to a graph that has an f-regular partition.

THEOREM 1 (see [11], [9]). *For every m, $\epsilon > 0$ and nonincreasing function $f : \mathbb{N} \to (0, 1)$, there is an integer $T = T_1(f, \epsilon, m)$ so that given a graph G with at least T vertices, one can add-to/remove-from G at most ϵn^2 edges and thus get a graph G' that has an f-regular partition of order k for some k with $m \le k \le T$.*

Our first result in this paper is a new short proof of the above theorem. The proof is a simple application of the variant of the regularity lemma of [2] mentioned above. Basing the proof on this method provides both explicit bounds and a polynomial time algorithm for finding the partition and the necessary modifications. Section 2 consists of the proof of Theorem 1, and in section 3 we describe a deterministic polynomial time algorithm for finding a regular partition and a set of modifications that are guaranteed by this theorem.

We now turn to the second result of this paper. In many cases, one applies the regularity lemma on a graph G to get an ϵ-regular partition $\mathcal{V} = \{V_i : 1 \le i \le k\}$

[1] This is not accurate because Definition 1.4 requires *all* pairs to be $f(k)$-regular, while Lemma 1.3 guarantees that only *most* pairs are regular.

and then defines a weighted complete graph on k vertices $\{1,\ldots,k\}$, in which the weight of the edge connecting vertices (i,j) is $d(V_i,V_j)$. This relatively small weighted graph, sometimes called the *regularity-graph* (or *reduced graph*) of G, carries a lot of information on G. For example, it can be used to approximately count the number of copies of any fixed small graph in G, and to approximate the size of the maximum-cut of G. A natural question, which was suggested to us by Sudan [13], is how different can two regularity-graphs of the same graph be? We turn to define what it means for two regularity graphs, or equivalently for two regular partitions, to be ϵ-isomorphic.

DEFINITION 1.5 (ϵ-isomorphic). *We say that two partitions $\mathcal{U}=\{U_i : 1\le i\le k\}$ and $\mathcal{V}=\{V_i : 1\le i\le k\}$ of a graph G are ϵ-isomorphic if there is a permutation $\sigma:[k]\to[k]$, such that for all but at most $\epsilon\binom{k}{2}$ pairs $1\le i<j\le k$, we have $|d(U_i,U_j)-d(V_{\sigma(i)},V_{\sigma(j)})|\le\epsilon$.*

We first show that if one considers the standard notion of an ϵ-regular partitions (as in Definition 1.2), then ϵ-regular partitions of the same graph are not necessarily similar. In fact, as the following theorem shows, even $f(k)$-regular partitions of the same graph, where $f(k)=1/k^{\delta}$, are not necessarily similar. A variant of this theorem has been proved by Lovász [9].

THEOREM 2. *Let $f(k)=1/k^{1/4}$. For infinitely many k and for every $n>n_2(k)$ there is a graph $G=(V,E)$ on n vertices with two f-regular partitions of order k that are not $\frac{1}{4}$-isomorphic.*

The proof of Theorem 2 provides explicit examples. We note that an inexplicit probabilistic proof shows that the assertion of the theorem holds even for $f(k)=\Theta(\frac{\log^{1/3}k}{k^{1/3}})$. See section 4 for more details.

Using the terminology of Definition 1.2, the above theorem and its proof can be restated as saying that for any (small) $\epsilon>0$ and all large enough $n>n_0(\epsilon)$, there exists an n vertex graph that has two ϵ-regular partitions of order ϵ^{-4}, which are not $\frac{1}{4}$-similar. Therefore, ϵ-regular partitions of the same graph may be very far from isomorphic.

Recall now that Theorem 1 guarantees that for any function f, any graph can be slightly modified in a way that the new graph admits an f-regular partition. As the following theorem shows, whenever $f(k)<1/2k^2$ all of the regular partitions of the new graph must be close to isomorphic.

THEOREM 3. *Let $f(k)$ be any function satisfying $f(k)\le\min\{1/2k^2,\frac{1}{8}\epsilon\}$, and suppose $\mathcal{U}$ and $\mathcal{V}$ are two f-regular partitions of some graph G on $n\ge\frac{8k}{\epsilon}$ vertices. Then $\mathcal{U}$ and $\mathcal{V}$ are ϵ-isomorphic.*

This theorem illustrates the power of f-regular partitions, showing that (for $f(k)<1/2k^2$) they enjoy properties that do not hold for usual regular partitions. Observe that the above results imply that when, e.g., $f(k)=\omega(\frac{\log^{1/3}k}{k^{1/3}})$, then two f-regular partitions of the same graph are not necessarily similar, whereas whenever $f(k)<1/2k^2$ they are. It may be interesting to find a tight threshold for f that guarantees ϵ-isomorphism between f-regular partitions of the same graph. It should also be interesting to find a similar threshold assuring that partitions of two *close* graphs are similar.

2. Proof of Theorem 1. In this section we show how to derive Theorem 1 from a variant of the regularity lemma due to Alon et al. [2]. Before we get to the proof we observe the following three simple facts. First, a standard probabilistic argument shows that for every δ and η, and for every large enough $n>n_0(\delta)$, there exists a

δ-regular pair (A,B) with $|A|=|B|=n$ and $d(A,B)=\eta$.[2] The additional two facts we need are given in the following two claims, where we use the notation $x=y\pm\epsilon$ to denote the fact that $y-\epsilon\le x\le y+\epsilon$.

CLAIM 2.1. *Let δ and γ be fixed positive reals, and let $n>n_0(\delta,\gamma)$ be a large enough integer. Suppose (A,B) is a δ-regular pair satisfying $d(A,B)=\eta\pm\gamma$ and $|A|=|B|=n$. Then, one can add or remove at most $2\gamma n^2$ edges from (A,B) and thus turn it into a 3δ-regular pair satisfying $d(A,B)=\eta\pm\delta$.*

Proof. Let us assume that $d(A,B)=\eta+\gamma$. The general case where $\eta-\gamma\le d(A,B)\le\eta+\gamma$ is similar. Suppose we delete each of the edges connecting A and B with probability $\frac{\gamma}{\eta+\gamma}$. Clearly the expected value of $d(A,B)$ after these modifications is η and, assuming n is large enough, we get from a standard application of Chernoff's bound that the probability that the new density deviates from η by more than δ is at most $\frac{1}{4}$. Also, the expected number of edges removed is γn^2 and again, if n is large enough, the probability that we removed more than $2\gamma n^2$ edges is at most $\frac{1}{4}$. Consider now two subsets $A'\subseteq A$ and $B'\subset B$, each of size δn. As (A,B) was initially δ-regular, we initially had $d(A',B')=(\eta+\gamma)\pm\delta$. As each edge is removed with a probability of $\frac{\gamma}{\eta+\gamma}$, the expected value of $d(A',B')$ after these modifications is $\eta\pm\frac{\delta\eta}{\eta+\gamma}=\eta\pm\delta$. By Chernoff's bound, we get that for large enough n for every such pair (A',B') the probability that $d(A',B')$ deviates from $\eta\pm\delta$ by more than δ is bounded by 2^{-4n}. As there are less than 2^{2n} choices for (A',B'), we get that with a probability of at least $\frac{3}{4}$ all pairs (A',B') have density $\eta\pm2\delta$. To recap, we get that with a probability of at least $\frac{1}{4}$ we made at most $2\gamma n^2$ modifications, $d(A,B)=\eta\pm\delta$ and $d(A',B')=\eta\pm2\delta$, implying that (A,B) is 3δ-regular. ∎

CLAIM 2.2. *Let (A,B) be a pair of vertex sets with $|A|=|B|=n$. Suppose A and B are partitioned into subsets $A_1,\dots,A_l$ and $B_1,\dots,B_l$ such that all pairs (A_i,B_j) are $\frac{1}{4}\delta^2$-regular and satisfy $d(A_i,B_j)=d(A,B)\pm\frac{1}{4}\delta$. Then (A,B) is δ-regular.*

Proof. Assume, towards a contradiction, that there are two subsets $A'\subseteq A$ and $B'\subseteq B$ of size at least δn each, such that $|d(A',B')-d(A,B)|>\delta$. By averaging, we may assume, without loss of generality, that $|A'|=|B'|=\delta n$.

Set $A'_i=A'\cap A_i$ and $B'_i=B'\cap B_i$. The number of pairs $(a\in A',b\in B')$, where $a\in A'_i$, $b\in B'_j$, and either $|B'_j|<\frac{1}{4}\delta^2|B_j|$ or $|A'_i|<\frac{1}{4}\delta^2|A_i|$ is bounded by $\frac{1}{2}\delta^3n^2$. Therefore, the possible contribution of such pairs to $d(A',B')$ is bounded by $\frac{1}{2}\delta$.

Consider now the pairs (A'_i,B'_j) satisfying $|B'_j|\ge\frac{1}{4}\delta^2|B_j|$ and $|A'_i|\ge\frac{1}{4}\delta^2|A_i|$. As (A_i,B_j) is $\frac{1}{4}\delta^2$-regular we have $d(A'_i,B'_j)=d(A_i,B_j)\pm\frac{1}{4}\delta$. As $d(A_i,B_j)=d(A,B)\pm\frac{1}{4}\delta$, we conclude that $d(A'_i,B'_j)=d(A,B)\pm\frac{1}{2}\delta$. As the pairs discussed in the preceding paragraph can change $d(A',B')$ by at most $\frac{1}{2}\delta$, we conclude that $d(A',B')=d(A,B)\pm\delta$, showing a contradiction as needed. ∎

The following is the strengthened version of the regularity lemma, due to Alon et al. [2], from which we will deduce Theorem 1.

LEMMA 2.3 (see [2]). *For every integer m and function $f:\mathbb{N}\to(0,1)$ there exists an integer $T=T_{2.3}(m,f)$ with the following property: If G is a graph with $n\ge T$ vertices, then there exists a partition $\mathcal{A}=\{V_i\ :\ 1\le i\le k\}$ and a refinement $\mathcal{B}=\{V_{i,j}\ :\ 1\le i\le k,1\le j\le l\}$ of $\mathcal{A}$ that satisfy the following:*

1. *$|\mathcal{A}|=k\ge m$ but $|\mathcal{B}|=kl\le T$.*
2. *For all $1\le i<i'\le k$ and for all $1\le j,j'\le l$, but at most $f(k)l^2$ of them, the pair $(V_{i,j},V_{i',j'})$ is $f(k)$-regular.*

[2]Here and throughout the rest of this paper, we say that $d(A,B)=\eta$ if $|e(A,B)-\eta|A||B|\,|\le1$. This avoids rounding problems arising from the fact that $\eta|A||B|$ may be nonintegral.

3. *All* $1 \le i < i' \le k$*, but at most* $f(0)\binom{k}{2}$ *of them, are such that for all* $1 \le j, j' \le l$*, but at most* $f(0)l^2$ *of them,* $|d(V_i, V_{i'}) - d(V_{i,j}, V_{i',j'})| < f(0)$ *holds.*

Proof of Theorem 1. Given a graph G, an integer m, a real ϵ, and some function $f : N \mapsto (0,1)$ as an input to Theorem 1, let us apply Lemma 2.3 with the function $f'(k) = \min\{f^2(k)/12, \epsilon/8\}$ and with $m' = m$. By Lemma 2.3, if G has more than $T = T_{2.3}(m', f')$ vertices, then G has two partitions $\mathcal{A} = \{V_i \ : \ 1 \le i \le k\}$ and $\mathcal{B} = \{V_{i,j} \ : \ 1 \le i \le k, 1 \le j \le l\}$ satisfying the three assertions of the lemma. We claim that we can make less than ϵn^2 modifications in a way that all pairs (V_i, V_j) will become $f(k)$-regular.

We start by considering the pairs $(V_{i,j}, V_{i',j'})$, with $i < i'$, which are not $f'(k)$-regular. Every such pair is simply replaced by an $f'(k)$-regular bipartite graph of density $d(V_{i,j}, V_{i',j'})$. Such a pair exists by the discussion at the beginning of this section. The number of edge modifications needed for each such pair is at most $(n/kl)^2$, and by the second assertion of Lemma 2.3 we get that the total number of modifications we make at this stage over all pairs (V_i, V_j) is bounded by $\binom{k}{2} \cdot f'(k)l^2 \cdot (n/kl)^2 \le \frac{\epsilon}{8}n^2$.

We now consider the pairs $(V_i, V_{i'})$ that do not satisfy the third assertion of Lemma 2.3, that is, those for which there are more than $f'(0)l^2$ pairs $1 \le j, j' \le l$ satisfying $|d(V_i, V_{i'}) - d(V_{i,j}, V_{i',j'})| \ge f'(0)$. For every such pair $(V_i, V_{i'})$ we simply remove all edges connecting V_i and $V_{i'}$. As by the third assertion there are at most $f'(0)\binom{k}{2} < \frac{\epsilon}{8}k^2$ such pairs, the total number of edge modifications we make is bounded by $\frac{\epsilon}{8}n^2$.

We finally consider the pairs $(V_i, V_{i'})$ that satisfy the third assertion of Lemma 2.3. Let us denote $d = d(V_i, V_{i'})$. We start with pairs $(V_{i,j}, V_{i',j'})$ satisfying $|d - d(V_{i,j}, V_{i',j'})| \ge f'(0)$. Each such pair is replaced with an $f'(k)$-regular pair of density d. As there are at most $f'(0)l^2 \le \frac{\epsilon}{8}l^2$ such pairs in each pair (V_i, V_j), the total number of modifications made in the whole graph due to such pairs is bounded by $\frac{\epsilon}{8}n^2$. Let us now consider the pairs $(V_{i,j}, V_{i',j'})$ satisfying $|d - d(V_{i,j}, V_{i',j'})| \le f'(0)$. If $d(V_{i,j}, V_{i',j'}) = d \pm f'(k)$, then we do nothing. Otherwise, we apply Claim 2.1 on $(V_{i,j}, V_{i',j'})$ with $\eta = d$, $\gamma = |d - d(V_{i,j}, V_{i',j'})|$, and $\delta = f'(k)$. Note that here we are guaranteed to have $\gamma \le f'(0) \le \frac{1}{8}\epsilon$. Claim 2.1 guarantees that we can make at most $2\gamma(n/kl)^2 \le \frac{1}{4}\epsilon(n/kl)^2$ modifications and thus turn $(V_{i,j}, V_{i',j'})$ into a $3f'(k)$-regular pair with density $d \pm f'(k)$. The total number of modifications over the entire graph is bounded by $\frac{\epsilon}{4}n^2$.

To conclude, the overall number of modifications we have made in the above stages is less than ϵn^2, as needed. Moreover, at this stage all of the pairs $(V_{i,j}, V_{i',j'})$ satisfy $|d(V_{i,j}, V_{i',j'}) - d(V_i, V_{i'})| \le f'(k) \le \frac{1}{4}f(k)^2$ and they are all $\frac{1}{4}f^2(k)$-regular. Therefore, by Claim 2.2 all pairs (V_i, V_j) are $f(k)$-regular, as needed. ∎

3. Deterministic algorithmic version of Theorem 1. As mentioned before, we show that it is also possible to obtain an algorithmic version of Theorem 1. Here is a rough sketch, following the proof of Theorem 1 step by step. As described in [2], one can obtain the partition of Lemma 2.3 in polynomial time. In order to find the modifications that make it f-regular, the random graphs can be replaced by appropriate pseudorandom bipartite graphs. The last ingredient we need is an algorithm for finding the modifications to a bipartite graph (A, B) that are guaranteed by Claim 2.1. The algorithm we describe here combines the use of conditional probabilities (see, e.g., [3]) with a certain local condition that ensures regularity. We first describe such a condition.

Given a bipartite graph on a pair of vertex sets (A, B) we denote by $d_{C_4}(A, B)$ the density of four-cycles in (A, B), namely, the number of copies of C_4 divided by $\binom{|A|}{2}\binom{|B|}{2}$. A pair (A, B) is said to be ϵ-*quad-regular* if $d_{C_4}(A, B) = d^4(A, B) \pm \epsilon$. This local condition indeed ensures ϵ-regularity, as detailed in the following lemma. The proof of the lemma appears in [6] and is based on the results of [1].

LEMMA 3.1 (see [6]). *Let (A, B) be a bipartite graph on A and B where $|A| = |B| = n$ and $\delta > 0$. Then we have the following:*

1. *If (A, B) is $\frac{1}{4}\delta^{10}$-quad-regular, then it is δ-regular.*
2. *If (A, B) is δ-regular, then it is 8δ-quad-regular.*

We shall design a deterministic algorithm for the following slightly weaker version of Claim 2.1.

CLAIM 3.2. *There is a deterministic polynomial time algorithm that, given a $\frac{1}{200}\delta^{20}$-regular pair (A, B) with n vertices in each part (with n large enough) and $d(A, B) = \eta \pm \gamma$, modifies up to $2\gamma n^2$ edges and thus turns the bipartite graph into a 2δ-regular pair with edge density $d'(A, B) = \eta \pm \delta$.*

Note that the polynomial loss in the regularity measure with respect to Claim 2.1 can be evened by modifying the definition of f' in the proof of Theorem 1 so that $f'(k) = \min\{f^{40}(k)/1600, \epsilon/8\}$. Hence Claim 3.2 indeed implies an algorithm for finding the modifications and partition guaranteed by Theorem 1.

Proof of Claim 3.2. Assume $d(A, B) = \eta + \gamma$ and $\gamma > \delta$. The case $d(A, B) = \eta - \gamma$ can be treated similarly.

Consider an arbitrary ordering of the edges of (A, B) and a random process in which each edge is deleted independently with probability $\frac{\gamma}{\eta+\gamma}$. We first consider this setting and later show that a sequence of deterministic choices of the deletions can be applied so that the resulting graph satisfies the desired properties.

Define the indicator random variable X_i, $1 \le i \le t = \eta n^2$, for the event of *not* deleting the ith edge. Denote the number of four cycles in (A, B) by $s = d_{C_4}(A, B)\binom{n}{2}^2$ and arbitrarily index them by $1, \ldots, s$. For every C_4 in (A, B) define the indicator Y_i, $1 \le i \le s$, for the event of its survival (i.e., none of its edges being deleted). Also let $X = \sum_{i=1}^{t} X_i$ and $Y = \sum_{i=1}^{s} Y_i$, which account for the numbers of edges and four-cycles, respectively, at the end of this process. Now define the following conditional expectations for $i = 0, 1, \ldots, t$, where the expectation is taken over the random independent choice of X_i described above:

(1)
$$f_i(x_1, \ldots, x_i) = \mathbb{E}_{X_{i+1},\ldots,X_t}\left[n^4(X - \eta n^2)^2 + (Y - \eta^4\tbinom{n}{2}^2)^2 \mid X_1 = x_1, \ldots, X_i = x_i\right].$$

We first obtain an upper bound on f_0. Since $X \sim B((\eta + \gamma)n^2, \frac{\eta}{\eta+\gamma})$, hence $\mathbb{E}[(X - \eta n^2)^2] = V(X) = O(n^2)$ and thus the first term in the expression for f_0 is $O(n^6)$. The expectation of the second term is

$$\mathbb{E}[(Y - \eta^4\tbinom{n}{2}^2)^2] = \mathbb{E}[Y^2] - 2\mathbb{E}[Y]\eta^4\tbinom{n}{2}^2 + \eta^8\tbinom{n}{2}^4.$$

For the linear term we have $\mathbb{E}[Y] = \sum_{i=1}^{s}\mathbb{E}[Y_i] = s(\frac{\eta}{\eta+\gamma})^4$. As for the quadratic term, for any pair $1 \le i < j \le s$ of four-cycles which share no common edge, the corresponding Y_i and Y_j are independent and hence $\mathbb{E}[Y_iY_j] = (\frac{\eta}{\eta+\gamma})^8$. There are only $O(n^6)$ nondisjoint pairs of C_4s, thus $\mathbb{E}[Y^2] = \mathbb{E}[\sum_{1\le i,j\le s} Y_iY_j] = s^2(\frac{\eta}{\eta+\gamma})^8 \pm O(n^6)$. By Lemma 3.1, $d_{C_4}(A, B) = (\eta + \gamma)^4 \pm \frac{1}{25}\delta^{20}$ and so $s = ((\eta + \gamma)^4 \pm \frac{1}{25}\delta^{20})\binom{n}{2}^2$.

Therefore, we conclude that

$$\mathbb{E}[(Y-\eta^4\tbinom{n}{2}^2)^2] = s^2(\tfrac{\eta}{\eta+\gamma})^8 \pm O(n^6) - 2s(\tfrac{\eta}{\eta+\gamma})^4\eta^4\tbinom{n}{2}^2 + \eta^8\tbinom{n}{2}^4$$
$$\leq \tfrac{1}{5}\delta^{20}\tbinom{n}{2}^4 + O(n^6).$$

This implies that, altogether, for a large enough n, $f_0 \leq \frac{1}{4}\delta^{20}\binom{n}{2}^4$.

Each $f_i(x_1,\ldots,x_i)$ is a convex combination of $f_{i+1}(x_1,\ldots,x_i,0)$ and $f_{i+1}(x_1,\ldots,x_i,1)$. Thus, for some choice x_{i+1} for X_{i+1}, we get that $f_{i+1}(x_1,\ldots,x_{i+1}) \leq f_i(x_1,\ldots,x_i)$. Therefore, choosing an x_{i+1} that minimizes f_{i+1} sequentially for $i = 0,\ldots,t-1$ results in an assignment of $(x_1,\ldots,x_t)$ such that $f_t(x_1,\ldots,x_t) \leq f_0 \leq \frac{1}{4}\delta^{20}\binom{n}{2}^4$. In order to apply this process, one needs to be able to efficiently compute f_i. But this is straightforward, since for any partial assignment of values to the X_is, the mutual distribution of any pair Y_i, Y_j can be calculated in time $O(1)$. Therefore, since there are at most $O(n^8)$ pairs of four-cycles, computing the expected value of the sum in (1) requires $O(n^8)$ operations. Repeating this for each edge accumulates to $O(n^{10})$.[3]

To complete the proof of the claim, we only need to show that the modifications we obtained above, namely such that $(x_1,\ldots,x_t)$ satisfy $f_t(x_1,\ldots,x_t) \leq \frac{1}{4}\delta^{20}\binom{n}{2}^4$, are guaranteed to satisfy the conditions of the claim. Indeed, in this case, each of the two addends which sum up to f_t are bounded by $\frac{1}{4}\delta^{20}\binom{n}{2}^4$. By the first addend, the new edge density d' is $d' = \eta \pm \frac{1}{2}\delta^{10}$. Thus, with much room to spare, the conditions on the edge density and the number of modifications are fulfilled. Note that it also follows that $d'^4 = \eta^4 \pm 3\delta^{10}$ (e.g., whenever $\delta < \frac{1}{4}$), and the second addend implies that the new four-cycles density is $\eta^4 \pm \frac{1}{2}\delta^{10} = d'^4 \pm 4\delta^{10}$. By Lemma 3.1 the pair is now $4^{1/5}\delta$-regular, and hence the modified graph attains all of the desired properties. ∎

Remark. Another possible proof of Claim 3.2 can be obtained by using an appropriate eightwise independent space for finding $(x_1,\ldots,x_t)$ such that f_t attains at most its expected value.

4. Isomorphism of regular partitions. In this section we prove Theorems 2 and 3. In order to simplify the presentation, we omit all floor and ceiling signs whenever these are not crucial. We start with the proof of Theorem 2. The basic ingredient of the construction is a pseudorandom graph which satisfies the following conditions.

LEMMA 4.1. *Let k be a square of an odd prime power, then there exists a graph $F=(V,E)$ on $|V|=k$ vertices such that we have the following:*

1. *F is $\lfloor k/2 \rfloor$-regular, and hence $d(V,V) = \frac{\lfloor k/2 \rfloor}{k}$,*
2. *for any pair of vertex sets A and B, if $|A| \geq k^{\frac{3}{4}}$ and $|B| \geq k^{\frac{3}{4}}$, then $d(A,B) = d(V,V) \pm k^{-\frac{1}{4}}$.*

Proof. We use some known pseudorandom graphs as follows; see the survey [8] for further definitions and details. An (n,d,λ)-*graph* is a d-regular graph on n vertices, all of whose eigenvalues, except the first one, are at most λ in their absolute values. It is well known that if λ is much smaller than d, then such graphs have strong pseudorandom properties. In particular, (see, e.g., [3, Chapter 9]), in this case for any two sets of vertices A and B of G: $d(A,B) = \frac{d}{n} \pm \lambda(|A||B|)^{-\frac{1}{2}}$. Thus, it is easy to verify that a $(k, \lfloor \frac{k}{2} \rfloor, \sqrt{k})$-graph would satisfy the assertions of the lemma.

[3] Note that each edge affects only at most $O(n^6)$ pairs of four-cycles, thus the complexity can easily be reduced to $O(n^8)$, and, in fact, the complexity can be further reduced by a more careful implementation.

There are many known explicit constructions of (n, d, λ)-graphs. Specifically, we use the graph constructed by Delsarte and Goethals and by Turyn (see [8]). In this graph the vertex set $V(G)$ consists of all elements of the two-dimensional vector space over $GF(q)$, where q is a prime power, so G has $k = q^2$ vertices. To define the edges of G, we fix a set L of $\frac{q+1}{2}$ lines through the origin. Two vertices x and y of the graph G are adjacent if $x - y$ is parallel to a line in L. It is easy to check that this graph is $\frac{(q+1)(q-1)}{2} = \frac{q^2-1}{2}$ -regular. Moreover, because it is a strongly regular graph, one can compute its eigenvalues precisely and show that besides the first one they all are either $-\frac{q+1}{2}$ or $\frac{q-1}{2}$. Therefore, indeed, we obtain an $(k, \lfloor \frac{k}{2} \rfloor, \lambda)$-graph with $\lambda < \sqrt{k}$ as necessary. □

Proof of Theorem 2. We construct our example as follows. Pick a graph F on k vertices $V(F) = \{1, \ldots, k\}$ which satisfies the conditions of Lemma 4.1. Suppose $n \geq k^2$. The graph on n vertices G will be an $\frac{n}{k}$ blow-up of F: every vertex of F is replaced by an independent set of size $\frac{n}{k}$, and each edge is replaced by a complete bipartite graph connecting the corresponding independent sets. Every nonedge corresponds to an empty bipartite graph between the parts. Let $\mathcal{U} = \{U_i \ : \ 1 \leq i \leq k\}$ be the partition of $V(G)$ where U_i is an independent set which corresponds to the vertex i in F. It follows from the construction that for any $1 \leq i < j \leq k$ the edge density of (U_i, U_j) is either 0 or 1, and (U_i, U_j) is ϵ-regular for any $\epsilon > 0$. The second partition $\mathcal{V}$ is generated by arbitrarily splitting every U_i into k equal-sized sets $W_{i,t}$, $1 \leq t \leq k$, and setting $V_t = \bigcup_{i=1}^{k} W_{i,t}$. Note that for any $1 \leq i < j \leq k$ the edge density $d_G(V_i, V_j)$ is exactly $d_F(V(F), V(F))$. Yet by Lemma 4.1, $d_F(V(F), V(F)) = \frac{2e(F)}{k^2} = \frac{\lfloor k/2 \rfloor}{k}$, which for $k \geq 2$ is strictly between $\frac{1}{4}$ and $\frac{3}{4}$. Hence $\mathcal{U}$ and $\mathcal{V}$ are not $\frac{1}{4}$-similar, as $|d(U_i, U_j) - d(V_{i'}, V_{j'})| > \frac{1}{4}$ for all pairs $i < j$ and $i' < j'$.

Thus, we complete the proof of the theorem by showing that all pairs (V_i, V_j) are $k^{-\frac{1}{4}}$-regular. Suppose, towards a contradiction and without loss of generality, that there are subsets $A \subseteq V_1$ and $B \subseteq V_2$ such that $|A| \geq k^{-\frac{1}{4}}|V_1|$, $|B| \geq k^{-\frac{1}{4}}|V_2|$, and $|d(A, B) - d(V_1, V_2)| > k^{-\frac{1}{4}}$.

For any $1 \leq i \leq k$ we denote $A_i = A \cap W_{i,1}$ and $B_i = B \cap W_{i,2}$. For any vertex $x \in A$, let the *fractional degree* of x with respect to B be defined by $d_B(x) = e(\{x\}, B)/|B|$. Note that $d(A, B) = \frac{1}{|A|}\Sigma_{x \in A} d_B(x)$ and that if x_1 and x_2 come from the same $W_{i,1}$, then $d_B(x_1) = d_B(x_2)$. Therefore, $d(A, B)$ is a convex combination

$$d(A, B) = \sum_{i=1}^{k} \frac{|A_i|}{|A|} d_B(x_i)$$

of (at most) k possible fractional degrees of vertices in A, where for $1 \leq i \leq k$, x_i is an arbitrary member of $W_{i,1}$.

First, assume that $d(A, B) > d(V_1, V_2) + k^{-\frac{1}{4}}$. We sort the vertices of A by their fractional degrees with respect to B, and consider a subset $\hat{A}$ of V_1 which consists of the union of the $k^{\frac{3}{4}}$ sets $W_{i,1}$ which have the highest fractional degrees with respect to B. Since $|A| > k^{-\frac{1}{4}}|V_1| = |\hat{A}|$, it follows that $d(\hat{A}, B) \geq d(A, B)$. Similarly, by considering the fractional degrees of the vertices of B with respect to the new subset $\hat{A}$, we may obtain a subset $\hat{B}$ of V_2 such that $d(\hat{A}, \hat{B}) \geq d(\hat{A}, B) \geq d(A, B) > d(V_1, V_2) + k^{-\frac{1}{4}}$. It also follows that both $\hat{A}$ and $\hat{B}$ are unions of sets $W_{i,1}$ and $W_{i,2}$, respectively. Thus, the edge density $d(\hat{A}, \hat{B})$ is exactly the edge density of the corresponding vertex sets in F (both of size $k^{\frac{3}{4}}$). By Lemma 4.1, we get that $d(\hat{A}, \hat{B}) \leq d_F(V(F), V(F)) + k^{-\frac{1}{4}} = d_G(V_1, V_2) + k^{-\frac{1}{4}}$, which leads to a contradiction

and completes the proof of Theorem 2. The case where $d(A,B) < d(V_1,V_2) - k^{-\frac{1}{4}}$ can be treated similarly. □

Remark. By using the random graph $G(k,\frac{1}{2})$ one could establish an inexplicit probabilistic proof for an analog of Lemma 4.1. The proof applies standard Chernoff bounds on the number of edges between *any* pair of *small* vertex sets. This extends the result for any $k > 2$ and with a stronger regularity constraint. Repeating the proof of Theorem 2 with such a graph F implies that Theorem 2 holds even for $f(k) = \Theta(\frac{\log^{1/3} k}{k^{1/3}})$.

We conclude this section with the proof of Theorem 3.

Proof of Theorem 3. First assume that n, the number of vertices in the graph G, is divisible by k, and consider two f-regular partitions $\mathcal{U} = \{U_i \ : \ 1 \le i \le k\}$ and $\mathcal{V} = \{V_i \ : \ 1 \le i \le k\}$ of order k. Let $W_{i,j}$ denote $V_i \cap U_j$. Consider a matrix A where $A_{i,j} = \frac{|W_{i,j}|}{|V_i|}$ is the fraction of vertices of V_i in U_j, and note that A is doubly stochastic, that is, the sum of entries in each column and row is precisely 1. A well-known (and easy) theorem of Birkhoff [4] guarantees that A is a convex combination of (less than) k^2 permutation matrices. In other words, there are k^2 permutations $\sigma_1,\ldots,\sigma_{k^2}$ of the elements $\{1,\ldots,k\}$, and k^2 reals $0 \le \lambda_1,\ldots,\lambda_{k^2} \le 1$ such that $\sum_t \lambda_t = 1$ and $A = \sum_t \lambda_t A_{\sigma_t}$, where A_σ is the permutation matrix corresponding to σ. Let λ_p be the largest of these k^2 coefficients. Clearly, $\lambda_p \ge 1/k^2$, and observe that as A is a convex combination of the matrices A_{σ_t}, this means that for every $1 \le i \le k$ we have $|W_{i,\sigma_p(i)}| \ge \frac{1}{k^2}|V_i|$ and similarly $|W_{i,\sigma_p(i)}| \ge \frac{1}{k^2}|U_{\sigma_p(i)}|$. As both $\mathcal{V}$ and $\mathcal{U}$ are assumed to be $f(k)$-regular and $f(k) < \min\{1/k^2, \epsilon/4\}$, this guarantees that for all $1 \le i < j \le k$ we have

$$|d(V_i,V_j) - d(U_{\sigma_p(i)},U_{\sigma_p(j)})|$$

$$\le |d(V_i,V_j) - d(W_{i,\sigma_p(i)},W_{j,\sigma_p(j)})| + |d(W_{i,\sigma_p(i)},W_{j,\sigma_p(j)}) - d(U_{\sigma_p(i)},U_{\sigma_p(j)})| \le \frac{\epsilon}{2}\,,$$

completing the proof for this case.

We now justify our assumption that n is divisible by k: if this is not the case, we add $(n \mod k) < k$ isolated vertices to G and denote the new graph by G'. Now, consider partitions $\mathcal{U}'$ and $\mathcal{V}'$ of G', in which all sets have the same size $\lceil \frac{n}{k} \rceil$, by adding at most one isolated vertex to each cluster in $\mathcal{U}$ and $\mathcal{V}$. Since $\mathcal{U}$ and $\mathcal{V}$ are $f(k)$-regular, it is not difficult to verify that $\mathcal{U}'$ and $\mathcal{V}'$ are $\min\{1/k^2, \epsilon/4\}$-regular. Applying the above argument on G' with $\mathcal{U}'$ and $\mathcal{V}'$, we get that $\mathcal{U}'$ and $\mathcal{V}'$ are $\frac{1}{2}\epsilon$-isomorphic. However, for any $1 \le i < j \le k$ the edge densities $d(V_i,V_j)$ and $d(V_i',V_j')$ differ by at most $\frac{2}{|V_i'|} \le \frac{2k}{n} \le \frac{1}{4}\epsilon$, and the same holds for $\mathcal{U}$ and $\mathcal{U}'$. Therefore, we conclude that the partitions $\mathcal{U}$ and $\mathcal{V}$ of the original graph G are ϵ-isomorphic. □

Acknowledgments. We would like to thank Madhu Sudan for a conversation that initiated this study, and Laci Lovász for fruitful discussions. We would also like to thank the anonymous referee for helpful comments.

REFERENCES

[1] N. Alon, R. A. Duke, H. Lefmann, V. Rödl, and R. Yuster, *The algorithmic aspects of the regularity lemma*, J. Algorithms, 16 (1994), pp. 80–109.

[2] N. Alon, E. Fischer, M. Krivelevich, and M. Szegedy, *Efficient testing of large graphs*, Combinatorica, 20 (2000), pp. 451–476.

[3] N. Alon and J. H. Spencer, *The Probabilistic Method*, 2nd ed., John Wiley & Sons, New York, 2000.

[4] G. Birkhoff, *Three observations on linear algebra*, Univ. Nac. Tucumán. Rev. Ser. A, 5 (1946), pp. 147–151.
[5] T. Gowers, *Lower bounds of tower type for Szemerédi's uniformity lemma*, Geom. Funct. Anal., 7 (1997), pp. 322–337.
[6] Y. Kohayakawa, V. Rödl, and J. Skokan, *Hypergraphs, quasi-randomness, and conditions for regularity*, J. Combin. Theory Ser. A, 97 (2002), pp. 307–352.
[7] J. Komlós and M. Simonovits, *Szemerédi's regularity lemma and its applications in graph theory*, in Combinatorics, Paul Erdös is Eighty, Vol. II, D. Miklós, V. T. Sós, and T. Szönyi, eds., János Bolyai Math. Soc., Budapest, 1996, pp. 295–352.
[8] M. Krivelevich and B. Sudakov, *Pseudo-random graphs*, in More Sets, Graphs and Numbers, E. Györi, G. O. H. Katona, and L. Lovász, eds., Bolyai Soc. Math. Stud. 15, Springer, Berlin, 2006, pp. 199–262.
[9] L. Lovász, *private communication*, 2006.
[10] L. Lovász and B. Szegedy, *Szemerédi's lemma for the analyst*, Geom. Funct. Anal., 17 (2007), pp. 252–270.
[11] V. Rödl and M. Schacht, *Regular partitions of hypergraphs: Counting lemmas*, Combin. Probab. Comput., 16 (2007), pp. 887–901.
[12] V. Rödl and M. Schacht, *Regular partitions of hypergraphs: Regularity lemmas*, Combin. Probab. Comput., 16 (2007), pp. 833–885.
[13] M. Sudan, *private communication*, 2005.
[14] E. Szemerédi, *Regular partitions of graphs*, in Problèmes Combinatories et Théorie des Graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, France, 1976), J. C. Bermond, J. C. Fournier, M. Las Vergnas, and D. Sotteau, eds., Colloq. Internat. CNRS 260, CNRS, Paris, 1978, pp. 399–401.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 288–299

COLORING RANDOM INTERSECTION GRAPHS AND COMPLEX NETWORKS*

MICHAEL BEHRISCH†, ANUSCH TARAZ‡, AND MICHAEL UECKERDT†

Abstract. We study the evolution of the chromatic number of a random intersection graph and show that, in a certain range of parameters, these random graphs can be colored optimally with high probability using different greedy algorithms. Experiments on real network data confirm the positive theoretical predictions and suggest that heuristics for the clique and the chromatic number can work hand in hand proving mutual optimality.

Key words. coloring, intersection graph, random graph, complex network

AMS subject classifications. 05C15, 05C80, 05C85

DOI. 10.1137/050647153

1. Introduction and results.

The classical random graph model, introduced by Erdős and Rényi in the early 1960s, considers a fixed set of n vertices and edges that exist with a certain probability, $p = p(n)$, independently from each other. It was shown to be inappropriate for describing real-world networks because it lacks certain features of those such as a scale-free degree distribution and the emergence of local clusters. One of the underlying reasons that are responsible for this mismatch is precisely the independence of the edges, in other words the missing transitivity: if vertices x and y exhibit a relationship of some kind in a real-world network and so do vertices y and z, then this suggests a connection between vertices x and z, too.

Intersection graphs. Suppose that we have a vertex set V and another set W. An *intersection graph* is a graph with vertex set V, where we assign to each vertex v a subset $W_v \subseteq W$ and connect two vertices v, v' by an edge if and only if their assigned sets W_v and $W_{v'}$ have nonempty intersection.

We call the ground set W from which the assigned sets are chosen *universal feature set* and its elements *features*. If feature $w \in W_v$, then we say that feature w is *assigned* to vertex v or simply that v *has* w. The set W_v is called the *feature set* of v. For a specified $w \in W$, let V_w be the set of vertices v that have feature w. We call V_w a *feature clique*, since it obviously induces a clique in the intersection graph. As usual, $\Gamma(v)$ denotes the set of neighbors of v, i.e., the set of vertices in V that have features with v in common.

Well-studied examples for intersection graphs are interval graphs on the real line. In this paper, however, we will only consider finite sets. Obviously, every graph is an intersection graph (simply pick an individual feature assigned only to the two vertices of every edge), but the fewer features we have, the more apparent the structure of the shared features inside the graph becomes.

*Received by the editors December 8, 2005; accepted for publication (in revised form) August 18, 2008; published electronically January 7, 2009. This work was supported in part by the DFG research center MATHEON in Berlin.

http://www.siam.org/journals/sidma/23-1/64715.html

†Institut für Informatik, Humboldt-Universität zu Berlin, 10099 Berlin, Germany (behrisch@informatik.hu-berlin.de, ueckerdt@informatik.hu-berlin).

‡Zentrum Mathematik, Technische Universität München, 80290 München, Germany (taraz@ma.tum.de).

Random intersection graphs. A *random intersection graph* on n vertices with a universal feature set W of size m is a random graph with vertex set $[n]$ where each vertex gets assigned a random set of features by choosing each feature independently with probability p. A sample of this probability space is denoted by $G_{n,m,p}$. We consider now, and in the following, $m := n^\alpha$, and will usually distinguish two cases: $\alpha > 1$ and $0 < \alpha < 1$. If the probability of $G_{n,m,p}$ having a property $\mathcal{A}$ tends to 1 with n tending to infinity, then we say that $G_{n,m,p}$ has property $\mathcal{A}$ *asymptotically almost surely* (a.a.s.).

It is sometimes convenient to look at the random intersection graph as a random bipartite graph with bipartition (V, W) and edges occurring between the two classes independently with probability p. Such a graph will be called a *generator*.

Several aspects of random intersection graphs have been studied before. Karoński, Scheinerman, and Singer-Cohen [11] study subgraph appearance in this model. Fill, Scheinerman, and Singer-Cohen [5] investigate the equivalence of $G_{n,m,p}$ to $G_{n,p}$, and Stark [14] analyzes its vertex degree distribution. Behrisch and Taraz [3] show how to reconstruct the feature structure when only the random intersection graph is given as input. A study of the component evolution is given by Behrisch in [2]. Some results concerning connectivity and cliques can be found in Singer [13]. Extensions to the model are proposed by Godehardt and Jaworski in [7], who modify the distribution of the sizes of the feature cliques. The practical relevance of random intersection graphs is studied by Newman, Strogatz, and Watts in [12] and by Guillaume and Latapy in [9].

The aim of this paper is to investigate the evolution of the chromatic number of $G_{n,m,p}$. As usual, denote by $\chi(G)$ the chromatic number of G and by $\omega(G)$ the size of the largest clique in G. The computation of these two fundamental parameters is long known to be NP-hard. Our main results are that for a random intersection graph $G = G_{n,m,p}$, where m and p lie in a certain range. Asymptotically almost surely, $\chi(G)$ and $\omega(G)$ can be computed efficiently by simple coloring heuristics and actually coincide.

THEOREM 1. *Let $m := n^\alpha$ with $\alpha > 0$ fixed and $p \ll \sqrt{\frac{1}{nm}}$. Then $G_{n,m,p}$ can a.a.s. be colored optimally in linear time and $\chi(G_{n,m,p}) = \omega(G_{n,m,p})$.*

THEOREM 2. *Let $m := n^\alpha$ with $0 < \alpha < 1$ fixed and $p \ll \frac{1}{m \ln n}$. Then $G_{n,m,p}$ can a.a.s. be colored optimally in linear time. Moreover, for $np > \ln^4 n$ we have a.a.s.*

$$\chi(G_{n,m,p}) = \omega(G_{n,m,p}) = (1 + o(1))np.$$

Note that in principle one could also state in Theorem 1 that for $np > \ln^4 n$ we have a.a.s. $\chi(G_{n,m,p}) = \omega(G_{n,m,p}) = (1 + o(1))np$, but this is redundant since $np > \ln^4 n$ and $p \ll \sqrt{\frac{1}{nm}}$ together imply $\alpha < 1$ and thus the two theorems overlap in this case. Figure 1 gives an overview about the parameter ranges where our theorems apply together with some basic properties of random intersection graphs.

Applications. We have tested our coloring heuristics on real-world networks from application areas such as the internet, cooperation graphs, and protein databases. Although we cannot prove that those networks can be modelled well with random intersection graphs having parameters in the range covered by our theorems, the heuristics described could color those graphs optimally in many cases—see section 4 for details. Still the question of, *why* one should try to *color* complex networks remains. Of course, knowledge of the chromatic number gives important structural information of a general nature, but while, for instance, the clique number is practically

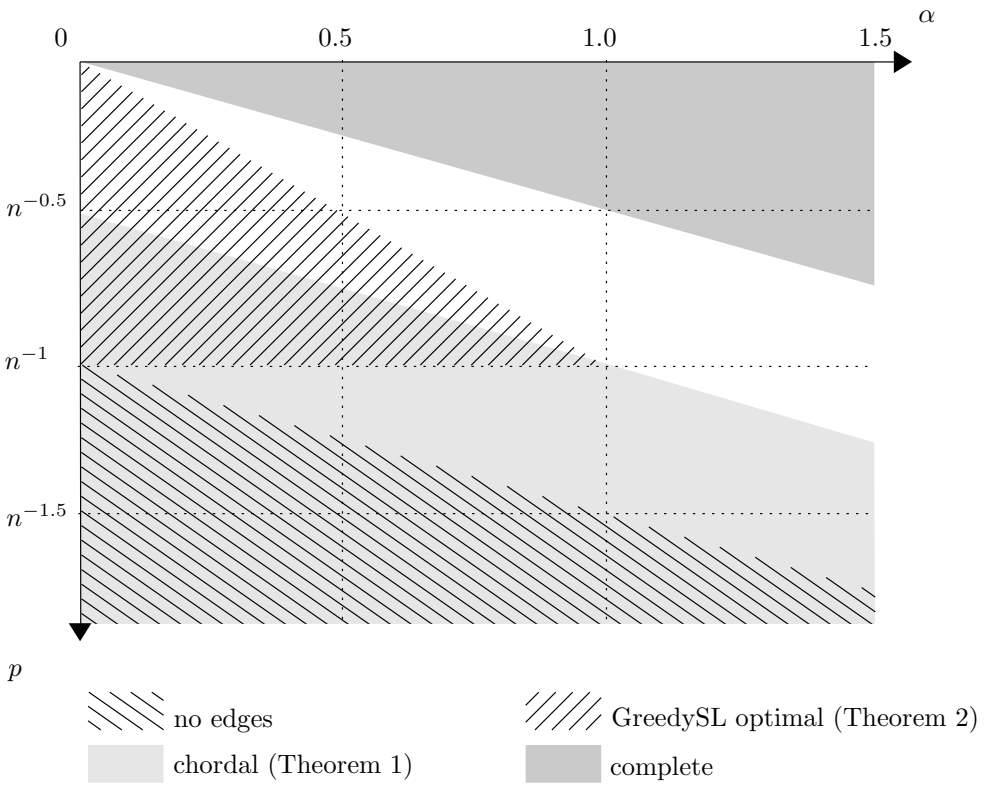


FIG. 1. *Ranges for p and α where we color optimally.*

meaningful—the size of the largest cluster in the network—the chromatic number seems to be of less immediate use.[1]

There is, however, one important application of the chromatic number, and this is exactly the clique number. Suppose we have a heuristic that tries to find the maximal size of a clique. If we also have a heuristic that tries to determine the minimum number of colors, and both of the proposed numbers coincide (or are at least very close to each other), then this proves that both numbers have already reached (near-) optimal values. This is precisely what we did in our experiments: we applied different heuristics discussed in an earlier paper [3] to find large cliques (and good clique covers) in the networks. At the same time, we tried to find good colorings of real-world networks using the greedy algorithms discussed in this paper. The results showed that, just as predicted for intersection graphs by Theorems 1 and 2, the proposed chromatic number and clique number indeed coincide in many cases.

In a way this is very reminiscent of the theory of perfect graphs. In fact, $G_{n,m,p}$ with m and p as in Theorem 1 is a.a.s. perfect, and we can thus use some of the perfect graph methodology to give a short proof of the theorem. For parameters m and p as in Theorem 2—although $\chi(G_{n,m,p}) = \omega(G_{n,m,p})$ a.a.s.—$G_{n,m,p}$ is not perfect and hence a different coloring strategy has to be used for this case.

[1] One possible application, not to be taken too seriously, could be to distribute film-stars to a minimum number of hotels (color classes) in such a way that costars of the same movie are not put in the same hotel, just to avoid trouble.

This paper is organized as follows. After a short section containing some auxiliary tools, we will prove Theorems 1 and 2 in sections 3.1 and 3.2, respectively. Our coloring experiments can be found in section 4, and a brief outlook concludes this paper.

2. Auxiliary lemmas. The following estimates are used without proof:

$$1 - ab \leq (1-a)^b \leq 1 - \frac{ab}{2} \qquad \text{for } 0 \leq a \leq 1, ab < 1. \tag{1}$$

Let X be a nonnegative random variable with expectation $\mu = \mathbb{E}[X]$. As a special case of Markov's inequality the first moment method states that

$$\mathbb{P}[X \geq 1] \leq \mu. \tag{2}$$

If X is binomially distributed random variable (n trials, each with probability p), then $\mu = np$, and we shall use the following variants of Chernoff's inequality (see section 2 in [10]):

$$\mathbb{P}[X \geq \mu + t] \leq \exp\left(-\frac{t^2}{2(\mu + t/3)}\right) \qquad \text{for } t \geq 0, \tag{3}$$

$$\mathbb{P}[X \leq \mu - t] \leq \exp\left(-\frac{t^2}{2\mu}\right) \qquad \text{for } t \geq 0, \tag{4}$$

$$\mathbb{P}[X \geq t] \leq \exp(-t) \qquad \text{for } t \geq 7\mu. \tag{5}$$

We first show that the probability that there is a feature clique in $G_{n,m,p}$ which deviates much from its expected size is exponentially small.

LEMMA 3. *Let $X_w := |V_w|$ be the random variable counting the number of vertices of a fixed feature w in a random intersection graph $G_{n,m,p}$ with $m := n^\alpha$ and $\alpha < 1$. Then*

$$\mathbb{P}\left[\exists w \in W : |X_w - pn| > (pn)^{\frac{3}{4}}\right] \leq m \exp\left(-\frac{(pn)^{\frac{1}{2}}}{3}\right).$$

Proof. The number of vertices chosen by a feature is a binomially distributed variable. Its deviation from its expected value can therefore be bounded by Chernoff inequalities (3) and (4). First, let w be fixed:

$$\mathbb{P}\left[X_w > pn + (pn)^{\frac{3}{4}}\right] \leq \exp\left(-\frac{(pn)^{\frac{3}{2}}}{2(pn + (pn)^{\frac{3}{4}}/3)}\right) \leq \frac{1}{2}\exp\left(-\frac{(pn)^{\frac{1}{2}}}{3}\right),$$

$$\mathbb{P}\left[X_w < pn - (pn)^{\frac{3}{4}}\right] \leq \exp\left(-\frac{(pn)^{\frac{3}{2}}}{2pn}\right) \leq \frac{1}{2}\exp\left(-\frac{(pn)^{\frac{1}{2}}}{3}\right).$$

By linearity of expectation (summing over all possible w) and Markov's inequality, the previous equation implies that

$$\mathbb{P}\left[\exists w \in W : |X_w - pn| > (pn)^{\frac{3}{4}}\right] \leq m \exp\left(-\frac{(pn)^{\frac{1}{2}}}{3}\right). \qquad \square$$

Since we are mostly interested in small feature sets, we need only an upper bound on their size.

LEMMA 4. *Let $X_v := |W_v|$ be the random variable counting the number of features for a fixed vertex v in a random intersection graph $G_{n,m,p}$ with $m := n^\alpha$ and $\alpha < 1$. Then, for $pm \leq 3 \ln n$*

$$\mathbb{P}\left[\exists v \in V : X_v > 21 \ln n\right] \leq \frac{1}{n^{20}}.$$

Proof. Very similar to the previous lemma, we have for a fixed vertex v, and for $pm \leq 3 \ln n$

$$\mathbb{P}\left[X_v > 21 \ln n\right] \overset{(5)}{\leq} \exp(-21 \ln n) = \frac{1}{n^{21}}.$$

Again summing over all vertices v yields the statement of the lemma. ∎

3. Proofs. In the following two subsections we describe two simple and well-known deterministic algorithms that find a proper coloring of a given input graph $G = (V, E)$ in linear time. Both algorithms are greedy heuristics: they color the vertices in a prescribed order and assign to each vertex the smallest color that has not been used for any of its neighbors which are already colored. Thus the main task is to prove the following: if the input graph G is a random intersection graph $G_{n,m,p}$ with parameters n, m, and p as given in Theorems 1 and 2, then these algorithms will a.a.s. produce a coloring with (at most) $\omega(G)$ different colors. Hence the coloring is optimal and $\chi(G) = \omega(G)$, as required.

The additional claim in Theorem 2, that a.a.s. $\omega(G)$ is of order np, will follow from the fact that the largest clique is a feature clique, which according to Lemma 3 is of that order.

3.1. Perfect elimination scheme. The aim of this subsection is to prove Theorem 1. Here is the basic idea of our coloring algorithm. We first try to order the vertices of the graph as $x_n, \ldots, x_1$ in such a way that for every vertex x_i the "remaining neighborhood" $\Gamma(x_i) \cap \{x_{i-1}, \ldots, x_1\}$ induces a clique in G. Having established this ordering, we greedily color the vertices in (reverse) order $x_1, \ldots, x_n$. Observe that this implies that vertices which are contained in many different cliques, e.g., those that have many features, will be colored relatively early.

Such an ordering is called a *perfect elimination scheme*, in short *PES*. Tarjan and Yannakakis [15] proved that, if a graph has a PES, then a so-called maximum cardinality search will produce a PES in linear time. If the graph doesn't have a PES, then the procedure returns an arbitrary ordering. This leads to the following greedy coloring heuristic:

ALGORITHM 1.

Input: *Graph $G = (V, E)$ on n vertices*
Output: *Coloring of G*
GREEDYCOLORPES*(G)*
(1) $A := \emptyset$
(2) ***for*** $i := 1$ ***to*** n
(3) *choose $x_i \in V \setminus A$ such that $|\Gamma(x_i) \cap A|$ is maximal*
(4) $A := A + x_i$
(5) ***for*** $i := 1$ ***to*** n
(6) *color x_i with the smallest color not occurring in $\Gamma(x_i)$*

The following three crucial facts have been known for a long time:

1. A graph G has a PES (and it can be found in linear time and as described previously) if and only if G is *chordal*, i.e., it does not contain an induced cycle with more than three vertices [15],
2. Chordal graphs are perfect [4, Chapter 5.5]; thus, in particular, $\chi(G) = \omega(G)$.
3. If a PES exists for G, then, using it as described previously, the greedy coloring procedure colors G optimally.

The last observation is a folklore result and obviously true: if the set of the already colored neighbors of every vertex x_i forms a clique when x_i is colored, then whenever a vertex x_i needs a new color k, we have just found a clique of size k, and hence k colors are really needed to color the graph.

Now all that remains is to prove that $G_{n,m,p}$ is chordal for the given parameters n, m, and p, which will be done in the following lemma.

LEMMA 5. *Let $m := n^\alpha$ for $\alpha > 0$ fixed and $p \ll \sqrt{\frac{1}{nm}}$. Then $G_{n,m,p}$ is a.a.s. chordal.*

Proof. Let $G = G_{n,m,p}$ be a random intersection graph and $B = (V \cup W, E_B)$ a bipartite generator of G. By definition, G is chordal if and only if it does not contain an induced cycle of length at least four. Suppose that $v_1, \ldots, v_k$ form an induced cycle C_k in G. Then there must exist features $w_1, \ldots, w_k$ such that w_i is a feature of both v_i and v_{i+1} for all $i \in [k-1]$, and w_k is a feature for both v_k and v_1. Moreover, all of the w_i are distinct, since otherwise the cycle wouldn't be induced. This yields a cycle $v_1, w_1, v_2, w_2, \ldots, v_k, w_k$ in the generator B. The probability for such a cycle in B can obviously be bounded from above by p^{2k}, and multiplying this with the number of possibilities to choose $v_1, \ldots, v_k$ and $w_1, \ldots, w_k$ we get

$$\mathbb{P}[G \text{ contains an induced } C_k] \leq n^k m^k p^{2k} = (nmp^2)^k.$$

The probability of G being not chordal is now bounded by

$$\begin{aligned}
\mathbb{P}[G \text{ is not chordal}] &\leq \sum_{k=4}^{min(n,m)} \mathbb{P}[G \text{ contains an induced } C_k] \\
&\leq \sum_{k=4}^{min(n,m)} (nmp^2)^k \\
&\leq \sum_{k=0}^{\infty} (nmp^2)^k - 1 = \frac{1}{1-nmp^2} - 1,
\end{aligned}$$

which tends to 0 for n tending to infinity because nmp^2 tends to 0. ∎

A second moment calculation (see Singer [13]) shows that $p = \sqrt{\frac{1}{nm}}$ is in fact the threshold function for the appearance of induced cycles of *fixed* length $k \geq 4$ in random intersection graphs. Thus for $p \gg \sqrt{\frac{1}{nm}}$ these graphs are a.a.s. not chordal.

3.2. Smallest last heuristic. The aim of this subsection is to prove Theorem 2. Again we employ a greedy strategy but this time the precomputed ordering $x_1, \ldots, x_n$ of the vertices is slightly different. Suppose we have already selected $x_n, \ldots, x_{i+1}$. Then among the remaining vertices x_i is the vertex with the smallest number of neighbors (among the remaining vertices). More precisely we have the following algorithm.

ALGORITHM 2.
Input: *Graph $G = (V, E)$ on n vertices*
Output: *Coloring of G*
GREEDYCOLORSMALLESTLAST*(G)*
(1) $A := V$
(2) ***for*** $i := n$ ***downto*** 1
(3) *choose $x_i \in A$ such that $|\Gamma(x_i) \cap A|$ is minimal*
(4) $A := A - x_i$
(5) ***for*** $i := 1$ ***to*** n
(6) *color x_i with the smallest color not occurring in $\Gamma(x_i)$*

As there may be more than one such ordering, we denote by $\chi_{\mathrm{SL}}(G)$ the maximum number of colors that GREEDYCOLORSMALLESTLAST(G) uses for an input graph G. It is well known [4, Chapter 5.2] that the number of colors used by the algorithm is always bounded from above by the maximal minimum degree of all subgraphs of G, plus one:

$$\chi_{\mathrm{SL}}(G) \leq 1 + \max_{H \subseteq G} \delta(H). \tag{6}$$

From this we derive the following simple proposition.

PROPOSITION 6. *If G is a graph such that*

(7) *every vertex v has less than $\omega(G)$ neighbors of degree at least $\omega(G)$,*

then

$$\chi_{SL}(G) = \omega(G) = \chi(G).$$

Proof. We claim that (7) implies that

$$1 + \max_{H \subseteq G} \delta(H) \leq \omega(G). \tag{8}$$

Suppose for a contradiction that there exists a subgraph H with $1 + \delta(H) > \omega(G)$. Let v be a vertex of minimal degree in H, i.e., $d_H(v) = \delta(H) \geq \omega(G)$. Then for *all* neighbors w of v in H we have

$$d_G(w) \geq d_H(w) \geq d_H(v) = \delta(H) \geq \omega(G),$$

and since there are $d_G(v) \geq d_H(v) = \delta(H) \geq \omega(G)$ neighbors of v in G, this contradicts the property in (7), which proves the claim in (8).

Now we are done, since

$$\chi(G) \leq \chi_{\mathrm{SL}}(G) \overset{(6)}{\leq} 1 + \max_{H \subseteq G} \delta(H) \overset{(8)}{\leq} \omega(G) \leq \chi(G). \qquad \square$$

Let us move back to intersection graphs. In the following we call a vertex v *rich* if it has at least two features. Obviously, the only way that a vertex can have degree at least $\omega(G)$ is if it is rich. Hence we have the following corollary.

COROLLARY 7. *Suppose that G is an intersection graph such that every vertex has less than $\omega(G)$ rich neighbors. Then*

$$\chi_{SL}(G) = \omega(G) = \chi(G). \qquad \square$$

In order to prove that in our random intersection graph the condition of the above corollary is a.a.s. satisfied, we first obtain an upper bound on the number of rich vertices in each feature clique.

LEMMA 8. *Let $m = n^\alpha$ for $0 < \alpha < 1$ fixed, $p \geq \frac{10 \ln^2 n}{n}$, and $t \geq 0$. Denote by ω_f the size of a largest feature clique in $G_{n,m,p}$. Then in a random intersection graph $G_{n,m,p}$, the probability that there exists a feature clique C with more than $\omega_f mp + t$ rich vertices is at most*

$$m \ \exp\left(-\frac{t^2}{2\omega_f mp + 2t/3}\right).$$

Proof. Let $C \subseteq V$ denote an arbitrary feature clique in G. For $v \in C$ we denote by $X_{C,v}$ the random variable which is 1 whenever v is rich and 0 otherwise. Then

$$\mathbb{P}[X_{C,v} = 1] = 1 - (1-p)^{m-1} \overset{(1)}{\leq} 1 - (1 - (m-1)p) \leq mp.$$

Let $X_C := \sum_{v \in C} X_{C,v}$ count the rich vertices in C. For the expectation of X_C we have

$$\mathbb{E}[X_C] = \sum_{v \in C} \mathbb{P}[X_{C,v} = 1] \leq \omega_f mp.$$

Using the Chernoff bound, we get

$$\begin{aligned}\mathbb{P}[X_C \geq \omega_f mp + t] &\leq \mathbb{P}[X_C \geq \mathbb{E}[X_C] + t] \\ &\overset{(3)}{\leq} \exp\left(-\frac{t^2}{2\mathbb{E}[X_C] + 2t/3}\right) \leq \exp\left(-\frac{t^2}{2\omega_f mp + 2t/3}\right).\end{aligned}$$

Of course the events $X_C \geq \omega_f mp + t$ are not independent of each other for overlapping feature cliques C, but, using linearity of expectation and the Markov inequality (2), we can bound the probability of existence of a feature clique with too many rich vertices by the expression in the lemma. ∎

Proof of Theorem 2. We want to apply Corollary 7 and, hence, need to show that in $G = G_{n,m,p}$ every vertex has less than $\omega(G)$ rich neighbors. Recall that $m := n^\alpha$ with $0 < \alpha < 1$ fixed and $p \ll \frac{1}{m \ln n}$. First, observe that we can assume that $pn > \ln^4 n$, since otherwise p would be so small that we could apply Theorem 1 instead. Set

$$t := \max(3 \ln n, \sqrt{nmp^2} \ln n),$$

and consider an arbitrary small $\varepsilon > 0$. We shall make use of the following two technical observations (involving t) that will be verified later:

$$21 \ln n((1+\varepsilon)nmp^2 + t) \leq (1-\varepsilon)np, \tag{9}$$

$$m \exp\left(-\frac{t^2}{2(1+\epsilon)nmp^2 + 2t/3}\right) \leq n^{\alpha - 1}. \tag{10}$$

Again denote by ω_f the size of a largest feature clique in $G = G_{n,m,p}$ and consider the following events that have already been discussed in Lemmas 3, 4, and 8, respectively:

$\mathcal{A}$: for all $w \in W : ||V_w| - pn| < \varepsilon pn$,

$\mathcal{B}$: for all $v \in V : |W_v| \leq 21 \ln n$,

$\mathcal{C}$: every feature clique C has at most $\omega_f mp + t$ rich vertices.

Let Y_v be the number of rich neighbors of a vertex v. Then Y_v is bounded from above by the number of feature cliques containing v, multiplied with the number of rich vertices per feature clique, and we can then compare this to the size of a feature clique, which is a lower bound for $\omega(G)$. So if all of the events $\mathcal{A}, \mathcal{B}, \mathcal{C}$ hold, then

$$Y_v \leq 21 \ln n \ ((1+\varepsilon)pn \ mp + t) \overset{(9)}{\leq} (1-\varepsilon)np \overset{(\mathcal{A})}{<} \omega_f - 1 < \omega(G), \tag{11}$$

which would immediately prove (most of) the statements in Theorem 2 because of Corollary 7. To prove that $\omega(G) = (1+o(1))np$, note that by the estimate in (11) there is no vertex v with $\omega_f - 1$ rich neighbors, and hence there exists no clique of size ω_f containing only rich vertices. In turn, this implies that $\omega(G) = \omega_f$, since a clique which is not (subset of) a feature clique contains only rich vertices, and we are done because $\omega_f = (1+o(1))np$ by property $\mathcal{A}$.

Let us complete the proof by showing that a.a.s. all of the events $\mathcal{A}, \mathcal{B}, \mathcal{C}$ hold. Obviously

$$\mathbb{P}\left[\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}\right] = 1 - \mathbb{P}\left[\bar{\mathcal{A}}\right] - \mathbb{P}\left[\mathcal{A} \cap \bar{B}\right] - \mathbb{P}\left[\mathcal{A} \cap \mathcal{B} \cap \bar{\mathcal{C}}\right] \geq 1 - \mathbb{P}\left[\bar{\mathcal{A}}\right] - \mathbb{P}\left[\bar{\mathcal{B}}\right] - \mathbb{P}\left[\mathcal{A} \cap \bar{\mathcal{C}}\right],$$

so it suffices to check that all of the probabilities $\mathbb{P}\left[\bar{\mathcal{A}}\right], \mathbb{P}\left[\bar{\mathcal{B}}\right], \mathbb{P}\left[\mathcal{A} \cap \bar{\mathcal{C}}\right]$ tend to zero. For the first two probabilities, this is immediately implied by Lemma 3 (which applies because of $m < n$ and $pn > \ln^4 n$) and Lemma 4, respectively. For the latter probability it follows from Lemma 8 and observing that

$$\mathbb{P}\left[\bar{\mathcal{A}} \cap \mathcal{C}\right] \leq m \ \exp\left(-\frac{t^2}{2(1+\varepsilon)pn \ mp + 2t/3}\right) \overset{(10)}{\leq} n^{\alpha - 1},$$

which does tend to zero, since $\alpha < 1$.

Thus all that remains is to check the two technical observations (9) and (10). Considering (9), we distinguish two cases. For $\sqrt{nmp^2} > 3$ we have

$$\begin{aligned} 21 \ln n((1+\varepsilon)nmp^2 + \sqrt{nmp^2} \ln n) &\leq 40nmp^2 \ln n + 21\sqrt{nmp^2} \ln^2 n \\ &= np(40mp \ln n + 21\sqrt{m/n} \ln^2 n), \end{aligned}$$

which is smaller than $(1-\varepsilon)np$ because of $mp \ll \frac{1}{\ln n}$ and $\alpha < 1$.

And for $\sqrt{nmp^2} \leq 3$

$$\begin{aligned} 21 \ln n((1+\varepsilon)nmp^2 + 3 \ln n) &\leq 40nmp^2 \ln n + 63 \ln^2 n \\ &\leq 360 \ln^3 n + 63 \ln^2 n, \end{aligned}$$

which is smaller than $(1-\varepsilon)np$ because of $\frac{\ln^3 n}{n} \ll p$.

Considering (10), we distinguish two cases again. For $\sqrt{nmp^2} > 3$ we have

$$\begin{aligned} m \exp\left(-\frac{nmp^2 \ln^2 n}{2(1+\epsilon)nmp^2 + \frac{2}{3}\sqrt{nmp^2} \ln n}\right) &\leq m \exp\left(-\frac{nmp^2 \ln^2 n}{nmp^2 \ln n}\right) \\ &= m \exp(-\ln n) = n^{\alpha - 1}, \end{aligned}$$

and for $\sqrt{nmp^2} \leq 3$

$$m \exp\left(-\frac{9 \ln^2 n}{2(1+\epsilon)nmp^2 + \frac{2}{3} 3 \ln n}\right) \leq m \exp\left(-\frac{9 \ln^2 n}{100 + 2 \ln n}\right)$$
$$\leq m \exp(-\ln n) = n^{\alpha - 1}. \qquad \square$$

4. Experiments. We have tested our algorithms on eight real-world networks from different application areas. The first five graphs are the same as in [9]. "Internet" describes part of the internet computer network, "Web" is the link graph of a complex website, "Authors" denotes a coauthoring graph, "Actors" denotes a costarship graph of actors as found in the internet movie database, and "Proteins" is an interaction graph of proteins. For details, see [9] and [1]. The "Mercator" graph is a graph of the internet at router level taken from [8]. Moreover, "DIP" stands for "Dictionary of Interfaces in Proteins" and is a similarity graph of protein parts (vertices are protein interfaces that are adjacent if they are similar) studied in [6]. "Drugs" is the result of a search for "relatives" of 13 substances in a database of 2000 drugs, where an edge connects a pair of drugs which are relatives to the same test substance. Details concerning this network are described in [16].

TABLE 1
Statistics on the performance of the algorithms on eight real-world networks.

	Internet	Web	Authors	Actors	Proteins
n	75885	325729	16400	392340	2113
$\|E\|$	357317	1090108	29552	15038083	2203
Greedyχ	22	155	11	294	6
GreedyPESχ	21	156	8	294	6
GreedySLχ	20	155	8	294	6
largest clique	20	155	8	294	6
core size	996	1367	0	2647	0

	Mercator	DIP	Drugs
n	284805	5119	2000
$\|E\|$	449246	14434	163969
Greedyχ	38	42	381
GreedyPESχ	33	42	381
GreedySLχ	33	42	381
largest clique	13	42	381
core size	1453	0	432

In Table 1 Greedyχ, GreedyPESχ, and GreedySLχ denote the number of colors needed by a greedy coloring procedure that colors the vertices in the natural order (in which they were read), in a PES ordering (cf. Algorithm 1), and in a smallest last ordering (cf. Algorithm 2), respectively. Table 1 also states the size of the largest clique we were able to find in the graphs using the clique cover algorithm described in [3]. Obviously, the difference between the proposed number of colors and the proposed size of a largest clique is an upper bound of the distance of either number to the optimal value.

The results show that the coloring algorithms seem to perform well on real-world graphs. In all but one case we were able to color the graph optimally using the heuristic described in Algorithm 2.

We also performed an additional test to obtain some indication as to how difficult it really is to optimally color these particular input graphs. For this, we determined

the so-called k-core by repeatedly removing all vertices with degree smaller than k, where we set k as the size of the largest known clique. If the k-core were very small or of a simple structure for which one could easily find a k-coloring, then it would be trivial to extend this coloring to a valid and thus optimal k-coloring of the whole graph by reattaching the vertices in reverse order. (Note that this procedure is essentially identical to Algorithm 2, except that it is forced to stop when it realizes that all remaining vertices $x \in A$ satisfy $|\Gamma(x) \cap A| \geq k$.) However, as shown in Table 1, in many cases the size of the k-core is substantially larger than that of the largest known clique.

Finally, we remark that the large difference between the proposed coloring number and the proposed clique number for the Mercator test set is not so much a failure of the coloring algorithms. Instead, it seems mainly due to the fact that the clique cover algorithm, originally designed in [3] with the aim to find a good clique *cover*, cannot find a large clique on this instance—a simple enumeration method applied to the 52-core of the graph ($k = 52$ gives the last nonempty k-core; it has 81 vertices) identified a clique of size 27.

5. Outlook. For the ranges not covered by Theorems 1 and 2, the chromatic number seems to be more difficult to estimate. From the aforementioned result by Singer [13], it is clear that those graphs are no longer chordal for $p \gg \sqrt{\frac{1}{nm}}$ while the results on the clique cover [3] suggest that the feature cliques stay the dominant structural element up to $p < \min\{\frac{1}{5}m^{-\frac{2}{3}}, \frac{n}{8m^2}\}$.

In higher ranges, the approximation of the chromatic number by the size of the largest feature clique will not be very good. Using a different approach [17], we tried to establish a better lower bound via the independence number. Since the chromatic number of any graph is at least as high as the number of vertices divided by the size of a largest independent set, we obtain a lower bound on the chromatic number which beats the size of the largest feature clique, as the following result shows.

THEOREM 9 (see [17]). *Let $\varepsilon > 0$ be fixed, and let $m := n^\alpha$ with $\alpha > 0$ fixed and $\frac{\ln n}{m} \ll p \ll \sqrt{\frac{\ln n}{m}}$. Then a.a.s. the random intersection graph $G_{n,m,p}$ has no independent set of size*

$$(2+\varepsilon)\frac{\ln n}{mp^2},$$

which implies that

$$\chi(G_{n,m,p}) \geq \frac{p^2 mn}{(2+\varepsilon)\ln n} \gg pn.$$

Lower bounds on the independence number (which match the upper bounds by a logarithmic factor) can also be found in [17].

REFERENCES

[1] R. ALBERT, H. JEONG, AND A.-L. BARABÁSI, *Database of Self-Organized Networks*, http://www.nd.edu/networks/database/index.html.

[2] M. BEHRISCH, *Component evolution in random intersection graphs*, Electron J. Combin., 14 (2007), pp. 12.

[3] M. BEHRISCH AND A. TARAZ, *Efficiently covering complex networks with cliques of similar vertices*, Theoret. Comput. Sci., 355 (2006), pp. 37–47.

[4] R. DIESTEL, *Graph Theory*, Springer-Verlag, New York, 1997.

[5] J. A. Fill, E. R. Scheinerman, and K. B. Singer-Cohen, *Random intersection graphs when $m = \omega(n)$: An equivalence theorem relating the evolution of the $G(n, m, p)$ and $G(n, p)$ models*, Random Structures Algorithms, 16 (2000), pp. 156–176.

[6] C. Frömmel, C. Gille, A. Goede, C. Gröpl, S. Hougardy, T. Nierhoff, R. Preissner, and M. Thimm, *Accelerating screening of 3D protein data with a graph theoretical approach*, Bioinformatics, 19 (2003), pp. 2442–2447.

[7] E. Godehardt and J. Jaworski, *Two models of random intersection graphs and their applications*, Electron. Notes Discrete Math. 10, Elsevier, Amsterdam, 2001.

[8] R. Govindan and H. Tangmunarunkit, *SCAN+Lucent Internet Map From the ISI*, http://www.isi.edu/div7/scan/mercator/maps.html, 1999.

[9] J.-L. Guillaume and M. Latapy, *Bipartite structure of all complex networks*, Inform. Process. Lett., 90 (2004), pp. 215–221.

[10] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs*, John Wiley & Sons, New York, 2000.

[11] M. Karoński, E. R. Scheinerman, and K. B. Singer-Cohen, *On random intersection graphs: The subgraph problem*, Combin. Probab. Comput., 8 (1999), pp. 131–159.

[12] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Random graphs with arbitrary degree distributions and their applications*, Phys. Rev. E, 64 (2001), 026118.

[13] K. B. Singer, *Random Intersection Graphs*, Ph.D. thesis, Johns Hopkins University, Baltimore, MD, 1995.

[14] D. Stark, *The vertex degree distribution of random intersection graphs*, Random Structures Algorithms, 24 (2004), pp. 249–258.

[15] 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.

[16] M. Thimm, A. Goede, S. Hougardy, and R. Preissner, *Comparison of 2D similarity and 3D superposition. Application to searching a conformational drug database*, J. Chem. Inf. Comp. Sci., 44 (2004), pp. 1816–1822.

[17] M. Ueckerdt, *Färben von zufälligen Schnittgraphen*, Diploma thesis, Humboldt-Universität zu Berlin, Berlin, Germany, 2005.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 300–318

CENTERS FOR RANDOM WALKS ON TREES*

ANDREW BEVERIDGE†

Abstract. We consider two distinct centers which arise in measuring how quickly a random walk on a tree mixes. Lovász and Winkler [*Efficient stopping rules for Markov chains*, in Proceedings of the 27th ACM Symposium on the Theory of Computing, 1995, pp. 76–82] point out that stopping rules which "look where they are going" (rather than simply walking a fixed number of steps) can achieve a desired distribution exactly and efficiently. Considering an optimal stopping rule that reflects some aspect of mixing, we can use the expected length of this rule as a mixing measure. On trees, a number of these mixing measures identify particular nodes with central properties. In this context, we study a variety of natural notions of centrality. Each of these criteria identifies the barycenter of the tree as the "average" center and the newly defined focus as the "extremal" center.

Key words. Markov chain, random walk, stopping rule, tree, barycenter

AMS subject classifications. 60510, 60G40, 05C05

DOI. 10.1137/070687402

1. Introduction.

What node is most central with respect to a random walk on a tree $G = (V, E)$? We define $P = \{p_{ij}\}$ to be the matrix of transition probabilities, so $p_{ij} = 1/d(i)$ if $ij \in E$ and $p_{ij} = 0$ otherwise. Let the hitting time $H(i, j)$ be the expected time for a random walk starting at node i to get to node j. A natural definition for centrality is to require that the target node j minimize this hitting time for an appropriately chosen starting node i. We consider two natural choices for this starting node. First, we identify the "average" center c by drawing i from the stationary distribution π:

$$\sum_{i \in V} \pi_i H(i, c) = \min_{j \in V} \sum_{i \in V} \pi_i H(i, j). \tag{1}$$

Next, we choose the worst possible starting node for each target j. Let j' be a j-*pessimal* node satisfying $H(j', j) = \max_{i \in V} H(i, j)$. A target node a achieving

$$H(a', a) = \min_{j \in V} H(j', j) = \min_{j \in V} \max_{i \in V} H(i, j) \tag{2}$$

is the "extremal" center of the tree.

There are two classical centers for trees. A node achieving $\min_{i \in V} \max_{j \in V} d(i, j)$, where $d(i, j)$ is the length of the unique path between i and j, is the *center* of the tree G (or bicenter if there are two adjacent nodes achieving this minimum). In other words, the distance to the furthest node from the center is minimal among all nodes of the tree G. This node does not appear to have any central properties with respect to random walks.

The *barycenter* is the node (or two adjacent nodes) achieving $\min_{i \in V} \sum_{j \in V} d(i, j)$. A barycenter minimizes the total distance to all other nodes. The following proposition reveals that the barycenter is the "average" center of the tree with respect to

*Received by the editors April 4, 2007; accepted for publication (in revised form) September 3, 2008; published electronically January 7, 2009.
http://www.siam.org/journals/sidma/23-1/68740.html
†Department of Mathematics and Computer Science, Macalester College, Saint Paul, MN 55105 (abeverid@macalester.edu). This author's research was supported in part by NSA Young Investigator Grant H98230-08-1-0064.

random walks. If u, v are adjacent nodes, then let $V_{u:v}$ be the set of nodes in the subtree rooted at u after the removal of the edge uv. This notation is meant to emphasize that nodes in $V_{u:v}$ are closer to u than to v. For any $S \subset V$ and any distribution τ, let $\tau(S) = \sum_{k \in S} \tau_k$. For an undirected graph, Coppersmith, Tetali, and Winkler [5] define the *central node* to be a node c for which $H(i,c) \leq H(c,i)$ for every node i. The following equivalence has been known to the authors of [5].

PROPOSITION 1. *Let $G = (V, E)$ be a tree. The following statements for a node c are equivalent.*

(a) *The node c is a barycenter of the tree.*
(b) *The node c satisfies $H(i,c) \leq H(c,i)$ for every node i.*
(c) $\sum_{k \in V} \pi_k H(k,c) = \min_{i \in V} \sum_{k \in V} \pi_k H(k,i)$.
(d) *For every node i adjacent to c, $\pi(V_{i:c}) = \sum_{k \in V_{i:c}} \pi_k \leq 1/2$.*

We introduce the term *focus* to denote the "extremal" center of the tree. There are two types: primary foci and secondary foci.

DEFINITION. Let G be a tree. If $a \in V$ satisfies $H(a', a) = \min_{j \in V} \max_{i \in V} H(i,j)$, then a is a *primary focus* of G. When all of the a-pessimal nodes are contained in a single subtree $G' \subset G \backslash \{a\}$, the unique a-neighbor b in G' is also a focus of G. If $H(b', b) = H(a', a)$, then b is a primary focus. If $H(b', b) < H(a', a)$, then b is a *secondary focus* of G.

PROPOSITION 2. *Every tree G either has one focus or has two adjacent foci.*

When G has a single focus, we say that G is *focal*. When G has two adjacent foci, we say that G is *bifocal*. A bifocal tree may have two primary foci or it may have one primary focus and one secondary focus. For a bifocal tree G with adjacent foci a, b, we will see that a has central properties for nodes in $V_{a:b}$ and that b has central properties for nodes in $V_{b:a}$.

We consider some examples of foci for trees. Let P_n denote the path of length n on vertices $v_0, v_1, \ldots, v_n$. It is well known that

$$H(v_i, v_j) = \begin{cases} j^2 - i^2 & i \leq j, \\ (n-j)^2 - (n-i)^2 & i > j. \end{cases} \tag{3}$$

A v_i-pessimal node must be the furthest leaf from v_i. The node that achieves $\min_{v_i \in P_n} H(v_i', v_i)$ is the center of the path, so the unique focus of P_{2k} is v_k and the foci of P_{2k+1} are v_k and v_{k+1}. In the latter case, both nodes are primary foci since $H(v_k', v_k) = H(v_{2k+1}, v_k) = (k+1)^2 = H(v_0, v_{k+1}) = H(v_{k+1}', v_{k+1})$.

Let $B_{r,s}$ denote the broom graph consisting of a star with r leaves $u_1, u_2, \ldots, u_r$ centered at node c, along with a path of length s on nodes $c = v_0, v_1, \ldots, v_s$. Simple calculations (using formula (11) in section 2) for $B_{4k,4k}$ show that node v_k is the primary focus with $H(v_k', v_k) = H(u_1, v_{k-1}) = 9k^2 + 1$ and that v_{k-1} is the secondary focus with $H(v_{k-1}', v_{k-1}) = H(v_{4k}, v_{k-1}) = (3k+1)^2$. Moreover, this broom graph shows that the center, barycenter, and foci of the tree are distinct notions. Indeed, additional calculations show that the nodes v_{2k}, v_{2k+1} are centers of $B_{4k,4k}$ and the barycenter is $c = v_0$. These three types of centers are pairwise separated by distance $\Theta(k)$.

Having defined the average and extremal centers, we consider a variety of criteria for centrality with respect to random walks on trees. The barycenter satisfies each "average" criterion and one of the foci of the tree (or both) satisfies each "extremal" criterion. Many of these criteria concern exact mixing measures defined via lengths of stopping rules.

Given an initial node i and a target distribution τ, we can follow an optimal stopping rule (see section 2 for a precise definition) to halt a random walk starting at i so that the distribution of the final node is exactly τ. Denote the expected length of this optimal rule by $H(i,\tau)$. A number of parameterless mixing measures defined via stopping rules have been introduced and studied in [1], [2], [3], [9], [10]. Among the most important measures are the mixing time $T_{\text{mix}} = \max_{i\in V} H(i,\pi)$ and the reset time $T_{\text{reset}} = \sum_{i\in V} \pi_i H(i,\pi)$. We interpret T_{mix} as the pessimal mixing time and T_{reset} as the average mixing time.

Since the barycenter is so closely related to average mixing, a natural question is how $H(c,\pi)$ compares with T_{reset}.

PROPOSITION 3. *$H(c,\pi) \le 2T_{\text{reset}}$, where c is a barycenter of the tree. This bound is tight for a star $K_{1,k}$ with $k \ge 2$.*

$H(c,\pi)$ may be considerably smaller than T_{reset}. Indeed, consider a rooted m-ary tree of depth r. By symmetry its root c is the unique barycenter and the unique focus.

THEOREM 4. *If G is an m-ary tree of depth r rooted at node c, then*

$$H(c,\pi) = \frac{(m+1)(m^r+1)}{(m-1)(m^r-1)}\, r - \frac{m^2+6m+1}{2(m-1)^2} \tag{4}$$

and

$$T_{\text{reset}} = \frac{2m^{r+1} - rm^2 - 2m + r}{(m-1)^2}. \tag{5}$$

Holding m fixed and letting $r \to \infty$, mixing from the root takes $\Theta(r)$ steps while average mixing takes $\Theta(m^{r-1})$ steps. The exact mixing result of (4) is complementary to the following approximate mixing result for m-ary trees due to Diaconis and Fill (Example 4.60 in [6]). If σ_t is the distribution achieved by walking t steps from the root c, then for fixed m, as $r \to \infty$, the total variational distance $\|\sigma_t - \pi\|$ becomes small after $\frac{m+1}{m-1}r + \alpha r^{1/2}$ steps for a large constant α.

We now turn our attention to the extremal center. Our analysis of the foci relies heavily on stopping rules. We provide a definition for the foci of a distribution τ (which is actually a generalization of the definition of the foci of a tree; see Proposition 6). For any target distribution τ we associate one node or two adjacent nodes that are central with respect to stopping rules from singleton distributions to τ.

DEFINITION. A node i is a *focus* of the distribution τ when $H(i,\tau) < 1 + \sum_{j\in V} p_{ij} H(j,\tau)$.

The left-hand side of this equation is the expected length of an optimal rule. The right-hand side of the equation is the expected length of the rule "take one step from i (according to the transition probabilities $P = \{p_{ij}\}$) and then follow an optimal rule starting from this neighbor node to τ." The τ-foci are the nodes for which this composite rule is not optimal.

For example, consider the path P_2 on nodes v_0, v_1, v_2 and take our target to be $\pi = (1/4, 1/2, 1/4)$. We first consider an optimal stopping rule from the center to π. Let $\Gamma_1(v_1,\pi)$ be the rule "with probability $1/2$ take one step to a random neighbor, otherwise stay put." The expected length $E(\Gamma_1(v_1,\pi)) = 1/2$ and (13) in the next section shows that this rule is an optimal rule from v_1 to π. Now consider the rule $\Gamma_2(v_1,\pi)$: "take one step and then follow an optimal rule to π." This rule is not optimal since $E(\Gamma_2(v_1,\pi)) > 1 > 1/2$.

Finally, consider the rule $\Gamma(v_0,\pi)$ from v_0 to π: "take one step and then follow the optimal rule $\Gamma_1(v_1,\pi)$." This turns out to be an optimal rule with $E(\Gamma(v_0,\pi)) = 3/2$.

The analogous rule from v_2 to π is also optimal. Therefore, according to the previous definition, the center v_1 is the unique focus for P_2. An analogous argument shows that the two internal nodes of P_4 are the foci for π. A similar phenomenon holds for an arbitrary tree.

THEOREM 5. *Every distribution τ on a tree has either one focus or two adjacent foci. If τ has a unique focus u, then $H(i,\tau) = H(i,u) + H(u,\tau)$ for all i. If τ has two foci u, v, then for $i \in V_{u:v}$, $H(i,\tau) = H(i,u) + H(u,\tau)$ and for $i \in V_{v:u}$, $H(i,\tau) = H(i,v) + H(v,\tau)$.*

The key observation of this theorem is that for any node i, the rule "walk from i to the nearest τ-focus and then follow an optimal rule from that focus to τ" is an optimal rule from i to τ. In other words, the foci of τ are central with respect to all walks from nodes to τ.

Naturally, the foci of the tree coincide with the foci of π.

PROPOSITION 6. *The foci of the tree G are the foci of the distribution π.*

This proposition shows that the focus of a distribution is indeed a generalization of the focus of the tree.

Another important mixing measure is the forget time $T_{\text{forget}} = \min_\tau \max_{i \in V} H(i,\tau)$, where the minimum is taken over all target distributions. We interpret this quantity as the minimum expected time to "forget" the node we started from by following an optimal rule to some distribution. In spite of its rather unorthodox definition, the forget time is intimately connected to the mixing time and the reset time. For an undirected graph, we have the nontrivial equality $T_{\text{reset}} = T_{\text{forget}}$ (see [10]), which are within a factor of 4 of T_{mix} (see [2]).

For any graph, Lovász and Winkler [10] show that there is a unique distribution μ achieving the forget time. This distribution μ is central in an extremal sense: μ minimizes the expected length of a rule starting from the worst possible node. For a tree, μ is concentrated on the foci of G:

PROPOSITION 7. *If the node a is the unique focus of $G = (V,E)$, then μ is the singleton distribution on the focus a. If the adjacent nodes a and b are the foci of G, then*

$$\mu_i = \begin{cases} \frac{1}{2|E|}(H(b',b) - H(a',b)), & i = a, \\ \frac{1}{2|E|}(H(a',a) - H(b',a)), & i = b, \\ 0 & \textit{otherwise,} \end{cases}$$

where $H(i',i) = \max_{j \in V} H(j,i)$.

Another mixing measure with central properties is $T_{\text{bestmix}} = \min_{i \in V} H(i,\pi)$. The node achieving T_{bestmix} is the best possible starting node for achieving the stationary distribution. This formulation is dual in some sense to that of (1). As expected, the foci of the tree are central for this extremal problem.

THEOREM 8. *The quantity $T_{\text{bestmix}} = \min_{i \in V} H(i,\pi)$ is achieved by a focus of the tree. Specifically, if $H(a',b) < H(b',a)$, then node a uniquely achieves T_{bestmix}; if $H(a',b) > H(b',a)$, then node b uniquely achieves T_{bestmix}; and if $H(a',b) = H(b',a)$, then T_{bestmix} is achieved by both a and b.*

The broom graphs $B_{2,2}$, $B_{2,3}$, and $B_{4,3}$ in Figure 1 show that all three possibilities do occur.

Consider another mixing measure similar to the forget time. The start-independent time of a distribution σ is $T_{\text{si}}(\sigma) = \min_\tau \sum_{i \in V} \sigma_i H(i,\tau)$, where the minimum is taken over all target distributions. For a walk started from σ, $T_{\text{si}}(\sigma)$ is the minimum expected time to obtain a sample (from some distribution) that is independent of the

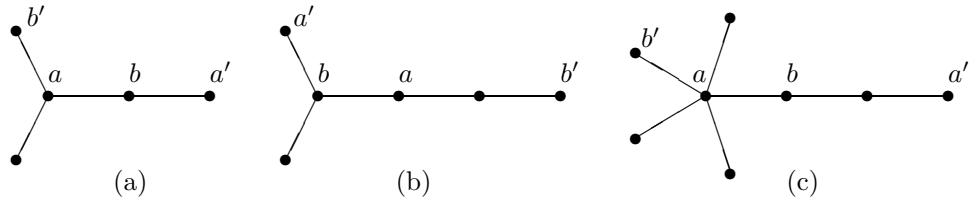


FIG. 1. *Trees with different foci achieving* T_{bestmix}. (a) $H(a', b) = H(b', a)$ *so* $T_{\text{bestmix}} = H(a, \pi) = H(b, \pi)$. (b) $H(a', b) < H(b', a)$ *so* $T_{\text{bestmix}} = H(a, \pi)$. (c) $H(a', b) > H(b', a)$ *so* $T_{\text{bestmix}} = H(b, \pi)$.

initial node of the walk (which was drawn from σ). We may interpret $T_{\text{si}}(\sigma)$ as the fastest way to "forget" that we started our walk from a node drawn from σ. A natural choice for our initial distribution is the stationary distribution. For a tree, the target distribution achieving $T_{\text{si}}(\pi)$ is central in an average sense and indeed the optimal target is concentrated on the barycenter.

PROPOSITION 9. *If c is a barycenter of the tree G, then*

$$T_{\text{si}}(\pi) = \min_{\tau} \sum_{k \in V} \pi_k H(k, \tau) = \sum_{k \in V} \pi_k H(k, c) = H(\pi, c).$$

We define the start-independent time of a graph to be $T_{\text{si}} = \max_{\sigma} T_{\text{si}}(\sigma)$, where the maximum is taken over all initial distributions. For a tree, the target distribution achieving T_{si} is central in an extremal sense, and indeed T_{si} can be achieved by taking either focus as a target. Furthermore, $T_{\text{si}} = T_{\text{forget}}$.

THEOREM 10. *For a tree G, we have $T_{\text{si}} = T_{\text{forget}}$. Moreover, if G has a unique focus a, then there exists a distribution ϕ such that $T_{\text{si}} = H(\phi, a)$. If G has two foci a and b, then there exists a distribution ϕ such that $T_{\text{si}} = H(\phi, a) = H(\phi, b)$.*

It is an open question as to how $T_{\text{si}}(\pi)$ and T_{si} compare to the other mixing measures for general graphs.

2. Preliminaries.

2.1. Random walks. Given an undirected, connected graph $G = (V, E)$, a random walk on G is a sequence of nodes $(w_0, w_1, \ldots, w_t, \ldots)$ such that the node w_t at time t is chosen uniformly from the neighbors of w_{t-1}. For nonbipartite G, as t tends to infinity the distribution of the node w_t tends to the so-called stationary distribution π, where $\pi_i = d(i)/2|E|$ and $d(i)$ is the degree of i. For bipartite G, we have convergence if we consider a "lazy walk" in which at each step we stay at the current node with probability $1/2$. For simplicity of exposition we will consider nonlazy walks (laziness simply doubles the expected length of our walks).

For two nodes i, j, the *hitting time* $H(i, j)$ is the expected length of a walk from i to j. The expected number of steps before a walk started at i returns to i is

$$\text{Ret}(i) = \frac{1}{\pi_i}. \tag{6}$$

For any undirected graph G, Coppersmith, Tetali, and Winkler [5] define a *central node* c of G to be a node which satisfies $H(i, c) \leq H(c, i)$ for all i. The existence of such a node for an undirected graph follows from their cycle reversing identity

$$H(i, j) + H(j, k) + H(k, i) = H(i, k) + H(k, j) + H(j, i). \tag{7}$$

The "random target identity" (cf. [8, equation (3.3)]) states that $\sum_{j\in V}\pi_j H(i,j)$ is independent of the initial node i. Multiplying (7) by π_k, summing over all k, and using this identity gives

$$\sum_{k\in V}\pi_k H(k,i)+H(i,j)=\sum_{k\in V}\pi_k H(k,j)+H(j,i) \tag{8}$$

for any nodes i,j.

2.2. Hitting times for trees. Random walks on trees have been extensively studied (cf. [4], [7], [11], [12]). We give a hitting time formula for trees that is equivalent to those found in [4], [7], and [11]. For adjacent nodes i and j,

$$H(i,j)=\sum_{k\in V_{i:j}}d(k)=2|E|\sum_{k\in V_{i:j}}\pi_k=2|E|\,\pi(V_{i:j}). \tag{9}$$

Indeed, let G' be the induced subtree on the nodes $V_{i:j}\cup\{j\}$. Let $d'(k)$ be the G'-degree of k and $H_{G'}(i,j)$ be the hitting time from i to j for this graph. By (6), $H(i,j)=H_{G'}(i,j)=\mathrm{Ret}_{G'}(j)-1=\sum_{k\in G'}d'(k)-1=\sum_{k\in V_{i:j}}d(k)$.

If i and j are neighbors, then (9) immediately gives

$$H(i,j)+H(j,i)=2|E|. \tag{10}$$

Furthermore, we can determine a hitting time formula for the general case. Define $\ell(i,k;j)=\frac{1}{2}(d(i,j)+d(k,j)-d(i,k))$ by the length of the intersection of the (i,j)-path and the (k,j)-path. This function is symmetric in i, and k and is zero if and only if $i=j$, $k=j$, or the nodes i and k are in different connected components of $G\setminus\{j\}$. Assume $d(i,j)=r$, and the (i,j)-path is given by $(i=i_0,i_1,i_2,\ldots,i_r=j)$. Using (9) and $\ell(i,k;j)$ yields

$$H(i,j)=\sum_{t=0}^{r-1}H(i_t,i_{t+1})=\sum_{k\in V}\ell(i,k;j)d(k). \tag{11}$$

Indeed, we can use formula (11) to recover formula (3) for hitting times on the path.

2.3. Stopping rules. We briefly summarize some results of Lovász and Winkler [9]. Let V^* be the space of finite walks on V, i.e., the set of finite strings $w=(w_0,w_1,w_2,\ldots,w_t)$, $w_i\in V$ and w_i adjacent to w_{i-1}. For a given initial distribution σ, the probability of w being the walk after t steps is

$$\Pr(w)=\sigma_{w_0}\prod_{i=0}^{t-1}p_{w_iw_{i+1}}.$$

A *stopping rule* Γ is a map from V^* to $[0,1]$ such that $\Gamma(w)$ is the probability of continuing given that w is the walk so far observed. We assume that with probability 1 the rule stops the walk in a finite number of steps.

Given another distribution τ on V, the *access time* $H(\sigma,\tau)$ is the minimum expected length of a stopping rule Γ that produces τ when started at σ. We say Γ is *optimal* if it achieves this minimum. For example, in the case that $\sigma=\tau$ are both singleton distributions on the node i, the rule "take no steps" is an optimal stopping rule with expected length 0, while the rule "walk until you return to i" is a nonoptimal stopping rule with expected length $\mathrm{Ret}(i)$.

Optimal stopping rules exist for any pair σ, τ of distributions, and the access time $H(\sigma, \tau)$ has many useful algebraic properties. When σ and τ are concentrated on nodes i and j, respectively (we write $\sigma = i, \tau = j$), then the access time $H(i, j)$ is the hitting time from i to j. Clearly, $H(\sigma, j) = \sum_{i \in V} \sigma_i H(i, j)$ and $H(\sigma, \tau) \leq \sum_{i \in V} \sigma_i H(i, \tau)$. The latter inequality is usually strict for nonsingleton distributions. For example, $0 = H(\pi, \pi) < \sum_{k \in V} \pi_k H(k, \pi) = T_{\text{reset}}$.

Given a stopping rule Γ from σ to τ, the *exit frequency* $x_i(\Gamma)$ is the expected number of times the walk leaves node i before halting. Exit frequencies partition the expected length of the walk: $E(\Gamma) = \sum_{k \in V} x_k(\Gamma)$. Exit frequencies are fundamental to virtually all access time results. A key observation, due to Pitman [13], is the "conservation equation"

$$\sum_{i \in V} p_{ij} x_i(\Gamma) - x_j(\Gamma) = \tau_j - \sigma_j. \tag{12}$$

It follows that the exit frequencies for two rules from σ to τ differ by $K\pi_i$ where K is the difference between the expected lengths of these rules. Hence the distributions σ and τ uniquely determine the exit frequencies for an optimal stopping rule between them, and we denote these optimal exit frequencies by $x_i(\sigma, \tau)$. Moreover,

$$\Gamma \text{ is an optimal stopping rule } \iff \exists k \in V,\ x_k(\Gamma) = 0. \tag{13}$$

Otherwise a rule with exit frequencies $x_k(\Gamma) - \pi_k \min_{i \in V}(x_i(\Gamma)/\pi_i)$ will have strictly smaller expected length while also satisfying (12). (See [9] for multiple ways to construct stopping rules from a given set of desired exit frequencies.) When $x_k(\Gamma) = 0$, we call the node k a (σ, τ)-*halting state*, or simply a *halting state* when the initial and target distributions are clear. The presence of a halting state is the single most useful criterion for determining whether a given rule is optimal. Note that an optimal rule may have multiple halting states, but we need only identify one such state to ensure that a rule is optimal.

Any three distributions σ, τ, and ρ satisfy the "triangle inequality"

$$H(\sigma, \rho) \leq H(\sigma, \tau) + H(\tau, \rho). \tag{14}$$

The right-hand side of this equation is the expected length of the composite rule that first follows an optimal stopping rule from σ to τ and then follows an optimal stopping rule from τ to ρ. The exit frequency for node k of this composite rule is $x_k(\sigma, \tau) + x_k(\tau, \rho)$. We have equality in (14) if and only if this composite rule is optimal. In particular, there must be some node k such that $x_k(\sigma, \tau) = 0$ and $x_k(\tau, \rho) = 0$. Considering the case where ρ is a singleton distribution, $H(\sigma, j) \leq H(\sigma, \tau) + H(\tau, j)$ for any node j and equality holds if and only if j is a halting state for an optimal rule from σ to τ. Hence

$$H(\sigma, \tau) = \max_{j \in V}(H(\sigma, j) - H(\tau, j)). \tag{15}$$

In the special case $\sigma = i$ and $\tau = \pi$, we have a particularly nice characterization due to the combination of (8) and (15):

$$j \text{ is an } (i, \pi)\text{-halting state} \iff H(j, i) = \max_{k \in V} H(k, i). \tag{16}$$

Let $j = i'$ denote such an *i-pessimal* node. We can reformulate this observation as

$$H(i, \pi) = H(i', i) - H(\pi, i). \tag{17}$$

2.3.1. Example: Mixing walks on P_3. We describe some optimal stopping rules from singleton distributions on $P_3 = (v_0, v_1, v_2, v_3)$ to $\pi = (1/6, 1/3, 1/3, 1/6)$. First, we construct an optimal mixing rule $\Gamma(v_0, \pi)$. By (13), a rule is optimal when it has a halting state. Equation (16) identifies v_3 as the unique halting state. Let $\Gamma(v_0, \pi)$ be the rule "choose a target node according to π and walk to that node." Since v_3 is never exited by this rule, $\Gamma(v_0, \pi)$ is optimal with expected length $H(v_0, \pi) = |\Gamma(v_0, \pi)| = \frac{1}{6}H(v_0, v_0) + \frac{1}{3}H(v_0, v_1) + \frac{1}{3}H(v_0, v_2) + \frac{1}{6}H(v_0, v_3) = 19/6$ by (3).

We now consider starting at the node v_1. Equation (13) again identifies v_3 as the unique halting state. For this starting node, choosing our target ahead of time does not result in an optimal rule: there is a nonzero chance of reaching v_3 before reaching v_0 (so v_3 would not be a halting state). Instead our heuristic is to try to stop as quickly as possible. The rule $\Gamma(v_1, \pi)$ is: "at $t = 0$, take a step with probability $2/3$ (and otherwise halt the walk for good). If the walk is still active at $t = 1$, then we are at either v_0 or v_2. If we are at v_2, then halt the walk. If we are at v_0, then stop with probability $1/2$, and otherwise keep walking until you reach v_3." Let us describe the behavior of this rule. At time $t = 0$, our distribution is $(0, 1, 0, 0)$. At time $t = 1$, our distribution is $(1/3, 1/3, 1/3, 0)$. Note that at time $t = 1$ our walk continues to be active only when we are at v_0. In this case we halt (with probability $1/2$) or continue walking (with probability $1/2$) until we reach v_3. When the rule finally terminates, our distribution is $(1/6, 1/3, 1/3, 1/6)$ and v_3 is a halting state. The expected length of this optimal rule is $H(v_1, \pi) = |\Gamma(v_1, \pi)| = \frac{2}{3} + \frac{1}{6}H(v_0, v_3) = 13/6$.

Finally, we consider another optimal (v_0, π)-rule. Let $\Gamma'(v_0, \pi)$ be the rule "take one step and then follow $\Gamma(v_1, \pi)$." Clearly, $|\Gamma'(v_0, \pi)| = 1 + |\Gamma(v_1, \pi)| = 19/6 = H(v_0, \pi)$, and indeed v_3 is a halting state for this composite rule. Interestingly, both of the rules $\Gamma(v_0, \pi)$ and $\Gamma(v_1, \pi)$ are optimal but are clearly distinct: $\Gamma'(v_0, \pi)$ always exits v_0 at $t = 0$ while $\Gamma(v_0, \pi)$ halts at $t = 0$ with probability $1/6$.

2.4. Mixing measures. Stopping rules provide a number of parameterless mixing measures. We define the *mixing time* T_{mix} to be the expected length of an optimal mixing rule starting from the worst initial node: $T_{\text{mix}} = \max_{i \in V} H(i, \pi)$. A node achieving this maximum is called *mixing pessimal*. The *forget time* T_{forget} is the smallest t such that there exists a distribution μ such that for every starting node, the expected time to attain μ via an optimal rule is at most t: $T_{\text{forget}} = \min_\tau \max_{i \in V} H(i, \tau)$. Theorem 4 (and the subsequent remark) in [10] establishes that the forget time is attained by a unique distribution given by

$$\mu_i = \pi_i \left(1 + \sum_{j \in V} p_{ij} H(j, \pi) - H(i, \pi) \right). \tag{18}$$

Furthermore, if a node is mixing pessimal, then it is also pessimal for μ, and every mixing pessimal node is a halting state for an optimal rule from μ to π.

The *reset time* $T_{\text{reset}} = \sum_{i \in V} \pi_i H(i, \pi)$ can be viewed as an average mixing time. Theorem 1 in [10] establishes the remarkable equality

$$T_{\text{forget}} = T_{\text{reset}} \tag{19}$$

for a random walk on an undirected graph. Moreover, for an undirected graph we have $T_{\text{reset}} \le T_{\text{mix}} \le 4T_{\text{reset}}$ (see [2, Corollary 8] and its subsequent remarks).

2.5. Start-independence. The following independence condition arises in applications of random walks. Let Γ be a stopping rule from σ to τ, and let $w_0, w_1, \ldots, w_T$

be a walk halted by Γ at time T. The *support* of σ, denoted S_σ, is the set of nodes i such that $\sigma_i > 0$. We associate a conditional distribution $\tau^{(i)}$ to each $i \in S_\sigma$ given by $\tau_k^{(i)} = \Pr\{w_T = k | w_0 = i\}$. In other words, $\tau_k^{(i)}$ is the probability that Γ stops the walk at k given that the walk started at i (which was drawn from σ). Clearly, $\sum_{i \in S_\sigma} \sigma_i \tau^{(i)} = \tau$, and we call the set $\{\tau^{(i)}\}_{i \in S_\sigma}$ the Γ-*decomposition* of τ.

The rule Γ is *start-independent* if $\tau^{(i)} = \tau$ for all $i \in S_\sigma$. The node at which a start-independent rule halts is independent of the initial node. Start-independent rules always exist: the rule "draw w_0 from σ and walk optimally from w_0 to τ" is a start-independent rule of expected length $\sum_{i \in V} \sigma_i H(i, \tau)$.

While start-independent rules are rarely optimal (for example, take $\sigma = \tau$), they arise naturally in applications requiring multiple independent samples from the stationary distribution of some state space. We obtain these samples by following an optimal mixing rule, accepting the current state, and then starting a new optimal mixing walk from this state. In this setting, T_{reset} is the expected length of a minimal start-independent rule from π to π. (See [3] for an extremal result concerning start-independent rules whose initial and target distributions are identical.)

We define the start-independent time of a distribution σ to be the minimum expected length of a start-independent rule with initial distribution σ:

$$T_{\text{si}}(\sigma) = \min_\tau \sum_{i \in V} \sigma_i H(i, \tau).$$

A quantity of natural interest is $T_{\text{si}}(\pi)$, the start-independent time for the stationary distribution. We would also like to determine the extremal behavior of $T_{\text{si}}(\sigma)$. The start-independent time of any singleton distribution is zero, so only the maximum case is nontrivial. We define the start-independent time of the graph to be

$$T_{\text{si}} = \max_\sigma T_{\text{si}}(\sigma) = \max_\sigma \min_\tau \sum_{i \in V} \sigma_i H(i, \tau).$$

3. The average and extremal centers. We begin with our characterization of the barycenter of the tree.

Proof of Proposition 1. The equivalence of (b) and (c) follows from (8): $H(i, c) \le H(c, i)$ for all i if and only if $\sum_{k \in V} \pi_k H(k, c) \le \sum_{k \in V} \pi_k H(k, i)$ for all i.

We show that (c) and (d) are equivalent. Assume $\pi(V_{i:c}) \le 1/2$ for every node i adjacent to c. For $j \in V_{i:c}$, $H(j, c) \le d(j, c) \sum_{k \in V_{i:c}} d(k) \le d(c, j) \sum_{k \in V_{c:j}} d(k) \le H(c, j)$ by (11); therefore c is the central node. Now assume that c is the central node and that $\pi(V_{i:c}) > 1/2$. Then $H(i, c) = \sum_{k \in V_{i:c}} d(k) > \sum_{k \in V_{c:i}} d(k) = H(c, i)$, a contradiction. If $\pi(V_{i:c}) = 1/2$ for some neighbor i of c, then i is also a central node.

Finally, we prove the equivalence of (a) and (d). For any adjacent nodes i and j, we have

$$\sum_{k \in V} d(k, i) = \sum_{k \in V} d(k, j) - |V_{i:j}| + |V_{j:i}| \tag{20}$$

and

$$\pi(V_{i:j}) = \sum_{k \in V_{i:j}} \pi_k = \frac{2|V_{i:j}| - 1}{2|E|}. \tag{21}$$

The node j has a neighbor i with $\pi(V_{i:j}) > 1/2 > \pi(V_{j:i})$ if and only if $|V_{i:j}| > |V_{j:i}|$ (by (21)) if and only if $\sum_{k \in V} d(k, i) < \sum_{k \in V} d(k, j)$ (by (20)) if and only if j is not the barycenter. ∎

Proof of Proposition 2. We divide the proof into two cases, depending on whether G has a focus a such that multiple subtrees of $G\backslash\{a\}$ contain an a-pessimal node.

Case 1. Let a' and a'' be a-pessimal nodes contained in different subtrees of $G\backslash\{a\}$. Any node $u \neq a$ is separated by a from at least one of a', a''. Without loss of generality, assume that a is on the unique (a', u)-path. Then $H(u', u) \geq H(a', u) = H(a', a) + H(a, u) > H(a', a)$ and therefore u is not a focus of the tree. Hence a is the unique focus of G.

Case 2. Suppose that a is a focus of G with all a-pessimal nodes in a single component of $G\backslash\{a\}$. Let b be the unique neighbor of a in this component. By definition, b is a focus of G and $H(b', b) \geq H(a', a)$. For any node $u \in V_{a:b}$, $H(u', u) \geq H(a', u) = H(a', a) + H(a, u) > H(a', a)$, so $u \in V_{a:b}$ is not a focus of G. The b-pessimal node b' must lie in $V_{a:b}$. Indeed, for any node $w \in V_{b:a}$, $H(w, b) < H(w, a) \leq H(a', a) \leq H(b', b)$. Similarly $a' \in V_{b:a}$. Now, considering $v \in V_{b:a}\backslash\{b\}$ we have $H(v', v) \geq H(b', v) = H(b', b) + H(b, v) > H(b', b) \geq H(a', a)$; therefore v is not a focus of G. ∎

The following corollary is immediate from the proof.

COROLLARY 11. *If G is focal with unique focus a, then there are multiple subtrees of $G\backslash\{a\}$ containing a-pessimal nodes. If G is bifocal with foci a, b, then each a-pessimal node is contained in $V_{b:a}$ and each b-pessimal node is contained in $V_{a:b}$.*

The barycenter is the average center for random walks on trees, so it is natural to compare $H(c, \pi)$ and $T_{\text{reset}} = T_{\text{forget}}$. Mixing from the barycenter never takes more than twice as long as the average mixing time.

Proof of Proposition 3. Let u be the unique neighbor of c on the path from c to a c-pessimal node c'. By Proposition 1, $\pi(V_{c:u}) \geq 1/2$. For any node $i \in V_{c:u}$ we have $H(i, \pi) = H(i, c) + H(c, \pi)$. Indeed, $c' \in V_{u:c}$ will also be i-pessimal and the composite rule corresponding to the right-hand side preserves c' as a halting state. Therefore (16) guarantees that this rule is optimal. Therefore

$$T_{\text{reset}} \geq \sum_{i \in V_{c:u}} \pi_i H(i, \pi) = \sum_{i \in V_{c:u}} \pi_i \left(H(i, c) + H(c, \pi)\right) \geq \pi(V_{c:u}) H(c, \pi) \geq \frac{1}{2} H(c, \pi).$$

For a star, we have $H(c, \pi) = 1/2$ and $T_{\text{reset}} = T_{\text{forget}} = 1$, so this bound is tight. ∎

There are trees for which $H(c, \pi) > T_{\text{reset}}$. Consider a broom graph $B_{4k,4k}$ with path nodes $(c = v_0, v_1, \ldots, v_{4k})$. Some simple calculations using (11) show that v_k and v_{k-1} are the foci of $B_{4k,4k}$. The forget time (and hence the reset time by (19)) is $T_{\text{forget}} = H(v_{4k}, \mu) < H(v_{4k}, v_{k-1}) = (3k+1)^2$. Using (11) and (17), the expected time to mix from the barycenter is $H(c, \pi) = H(c, c') - H(\pi, c') = (64k^2 - 1)/6$ which is strictly greater than $(3k+1)^2$ for $k \geq 4$. Of course, $(64k^2 - 1)/6 < 2(3k)^2 = 2H(v_{4k}, v_k) < 2T_{\text{forget}}$ as stipulated by Proposition 3.

On the other hand, $H(c, \pi)$ may be markedly smaller than the forget time (and hence the reset time) of the tree. Consider an m-ary tree of depth r with root c. Of course c is the center, the barycenter, and the focus of this tree. We adopt the following notation: $S_k = \{i | d(i, c) = k\}$ is the set of all nodes at level k. Let $c = i_0, i_1, \ldots, i_r$ be a path from c to a leaf i_r. The expected behavior of the walk at a node only depends on the level of the node, so we may use i_k as a representative for all nodes in S_k. A node is halting for this mixing walk if and only if it lies in S_r. We explicitly calculate $H(c, \pi) = H(i_r, c) - H(\pi, c)$ as per (17). Counting the degrees levelwise, the total

number of edges in an m-ary tree of depth r is

$$\frac{1}{2}\left(m+(m+1)\sum_{k=1}^{r-1}m^k+m^r\right)=\frac{m(m^r-1)}{m-1}.$$

Proof of Theorem 4. We start by showing that

$$H(i_r,i_{r-s})=\frac{2m^{s+1}-sm^2-2m+s}{(m-1)^2}. \tag{22}$$

Let G' be the connected component of $G\backslash\{i_{r-s}\}$ containing i_r. We partition $V(G')\cup\{i_{r-s}\}$ into sets $T_k=\{j|\ell(i_r,j;i_{r-s})=k\}$ so that

$$H(i_r,i_{r-s})=\sum_{k=1}^{s}k\sum_{j\in T_k}d(j)$$

by (11). We have $T_s=\{i_r\}$ and for $1\le k\le s-1$, T_k consists of the node i_{r-s+k} connected to $m-1$ copies of m-ary trees of depth $s-k-1$. Hence,

$$\sum_{j\in T_k}d(j)=(m+1)+(m-1)\left(1+\frac{2m(m^{s-k-1}-1)}{m-1}\right)=2m^{s-k}$$

for $1\le k\le s-1$ so that

$$\begin{aligned}H(i_r,i_{r-s})&=s+\sum_{k=1}^{s-1}k(2m^{s-k})=s+2\sum_{j=1}^{s-1}(s-j)m^j\\&=s+2s\sum_{j=1}^{s-1}m^j-2\sum_{j=1}^{s-1}jm^j=\frac{2m^{s+1}-sm^2-2m+s}{(m-1)^2}.\end{aligned}$$

By (17),

$$\begin{aligned}H(c,\pi)&=H(i_r,c)-H(\pi,c)=H(i_r,c)-\sum_{j\in V}\pi_jH(j,c)=H(i_r,c)-\sum_{k=0}^{r}\pi(S_k)H(i_k,c)\\&=\sum_{k=0}^{r}\pi(S_k)\left((H(i_r,i_k)+H(i_k,c))-H(i_k,c)\right)=\sum_{k=0}^{r}\pi(S_k)H(i_r,i_k)\\&=\frac{m-1}{2m(m^r-1)}\left(mH(i_r,i_0)+\sum_{k=1}^{r-1}m^k(m+1)H(i_r,i_{r-(r-k)})+m^rH(i_r,i_r)\right).\end{aligned}$$

Use formula (22) and simplify the result (we omit the details) to get the mixing result of (4).

We can now quickly determine the reset time of an m-ary tree. As per (19), $T_{\text{reset}}=T_{\text{forget}}$. By symmetry, the root is the unique focus of the m-ary tree and $T_{\text{forget}}=H(i_r,i_0)$. Formula (22) for $s=r$ gives (5). For completeness, we note that $T_{\text{mix}}=H(i_r,\pi)=H(i_r,i_0)+H(i_0,\pi)=T_{\text{forget}}+H(c,\pi)=T_{\text{reset}}+H(c,\pi)=\Theta(m^{r-1})$. ∎

4. The foci of a distribution. Recall that the node k is (i,τ)-halting when $x_k(i,\tau) = 0$. Two nodes i, j have a common halting state for τ when there exists a node k such that $x_k(i,\tau) = 0$ and $x_k(j,\tau) = 0$. A focus of a distribution τ on the tree G is a node u for which the rule "take one step from u and then follow an optimal rule from this random neighbor of u to τ" is not optimal, i.e., $H(u,\tau) < 1 + \sum_{i\in V} p_{ui} H(i,\tau)$. This is equivalent to saying that there is no node that is simultaneously τ-halting for u and all of its neighbors.

For example, the focus for the singleton distribution $\tau = u$ is the node u. Considering mixing walks, (16) states that k is a π-halting state for i if and only if $H(k,i) = \max_{j\in V} H(j,i) = H(i',i)$. Hence for a path of even length the unique center is the only π-focus, and for a path of odd length, the two central nodes are the π-foci. Also, the center of a star graph is the only π-focus.

We now prove Theorem 5, which states that for any tree G, every τ has one focus or two adjacent foci. Fixing τ, let i^* denote a halting state for an optimal stopping rule from i to τ.

LEMMA 12. *When i^* is an (i,τ)-halting state, then i^* is a (j,τ)-halting state whenever j and i^* are in different subtrees of $G\backslash\{i\}$.*

Proof. We are guaranteed that i is on the unique (j,i^*)-path. Consider the composite rule "walk from j until you reach i and then follow an optimal rule from i to τ." The kth exit frequency of this composite rule is $x_k(j,i)+x_k(i,\tau)$. In particular, $x_{i*}(j,i) + x_{i^*}(i,\tau) = 0$, therefore i^* is a halting state. By (13) this composite rule is therefore optimal: $H(j,\tau) = H(j,i) + H(i,\tau)$. ∎

LEMMA 13. *If $(j_1,\dots,i_1,i_2,\dots,j_2)$ is a path in the tree G, then the nodes j_1, j_2 cannot each be τ-halting states for both of the nodes i_1, i_2.*

Proof. Assume the conclusion is false. Equation (15) yields $H(i_k,\tau) = H(i_k,j_1) - H(\tau,j_1) = H(i_k,j_2) - H(\tau,j_2)$ for $k = 1,2$; therefore $-H(i_2,i_1) = H(i_1,j_1) - H(i_2,j_1) = H(i_1,\tau) - H(i_2,\tau) = H(i_1,j_2) - H(i_2,j_2) = H(i_1,i_2)$, a contradiction. ∎

Proof of Theorem 5. The case when τ is a singleton is trivial, therefore assume τ is not a singleton.

Case 1. There exists an edge uv such that u and v do not share a halting state for τ. Note that $u^* \in V_{v:u}$ and $v^* \in V_{u:v}$. Consider a set of optimal rules from the singletons to τ. Lemma 12 ensures that each node in $V_{u:v}$ has exactly the same τ-halting states and the nodes in $V_{v:u}$ all have the same τ-halting states. Therefore no node in $V\backslash\{u,v\}$ can be a focus for τ.

We claim that u is a focus of τ. Indeed, we show that the composite rule Γ from u to τ given by "take one step and then follow an optimal rule from that node to τ" is not optimal by proving that $x_k(\Gamma) > 0$ for all $k \in V$. Clearly u is not a halting state for Γ. If u is a leaf, then after our first step, we must be at v. Since every (v,τ)-halting state is contained in $V_{u:v} = \{u\}$, we must have $x_k(\Gamma) > 0$ for all $k \in V$. When u is not a leaf, let $i \in V_{u:v}$ be a neighbor of u. Then $x_k(\Gamma) \geq \frac{1}{d(u)}\left(x_k(v,\tau) + x_k(i,\tau)\right) > 0$. Indeed, after our first step from u, we are now at each of v, i with probability $1/d(u)$. Lemma 13 ensures that no node is simultaneously halting for both v and u. Lemma 12 states that i has the same halting states as u. Combining these observations yields $x_k(v,\tau) + x_k(i,\tau) > 0$ for every node k. Γ is not optimal by (13), and therefore u is a focus of τ. By a similar proof, v is also a focus of τ.

Case 2. Every neighboring pair of nodes shares a τ-halting state. Since τ is not a singleton, there exists a path of the form $(u^*,\dots,i,u,\dots,i^*)$ where u^* is a halting state for u but not for i, and u separates i from all of its τ-halting states. If u is

not a focus, then the neighbors of u have a common halting state j^*. Let $j \neq i$ be the neighbor of u on the (u, j^*)-path. The path $(u^*, \ldots, u, j, \ldots, j^*)$ is of the form forbidden by Claim 3, a contradiction. Therefore u must be a focus and unique. Indeed, u shares a τ-halting state with each of its neighbors, therefore u must have a halting state in at least two components of $G\backslash\{u\}$. If there was another focus v, then this would again imply the existence of a path forbidden by Lemma 13. □

Proof of Proposition 6. *Case* 1. G is bifocal with foci a, b. By Corollary 11, $a' \in V_{b:a}$, $b' \in V_{a:b}$, and by (16), $x_{a'}(a, \pi) = 0$ and $x_{b'}(b, \pi) = 0$. The argument is now identical to case 1 in the proof of Theorem 5 with $\tau = \pi, u = a, u^* = a', v = b, v^* = b'$.

Case 2. G is focal with focus a. Corollary 11 states that at least two subtrees of $G\backslash\{a\}$ contain a-pessimal nodes. By (16), these nodes are (a, π)-halting states. Lemma 12 now ensures that every neighboring pair of nodes share a π-halting state, therefore a is the only potential π-focus. From this point, the argument is identical to the end of case 2 in the proof of Theorem 5 with $\tau = \pi$ and $u = a$. □

Proof of Proposition 7. By (18), when i is not a focus of π we have $\mu_i = 0$, If G is focal, then μ is the singleton distribution on a. For G bifocal, rewrite (18) as

$$\mu_i = \pi_i \left(1 + \sum_{j \in V} p_{ij}(H(j, \pi) - H(i, \pi)) \right).$$

When $i \in V_{a:b}$ is a neighbor of a, Theorem 5 shows that $H(i, \pi) - H(a, \pi) = H(i, a)$. Equation (9) gives

$$\sum_{i \in V_{a:b}} p_{ai}(H(i, \pi) - H(a, \pi)) = \sum_{i \in V_{a:b}} p_{ai} H(i, a) = \frac{1}{d(a)}(H(a, b) - d(a)).$$

Considering the final neighbor b, (17) and (8) give

$$\begin{aligned} H(b, \pi) - H(a, \pi) &= H(b', b) - H(\pi, b) - H(a', a) + H(\pi, a) \\ &= H(b', a) + H(a, b) - H(a', b) - H(b, a) + H(\pi, a) - H(\pi, b) \\ &= H(b', a) - H(a', b). \end{aligned}$$

Thus our formula for μ_a becomes

$$\begin{aligned} \mu_a &= \frac{d(a)}{2|E|}\left[1 + \frac{1}{d(a)}\left(H(a, b) - d(a) + H(b', a) - H(a', b)\right)\right] \\ &= \frac{1}{2|E|}\left(H(a, b) + H(b', a) - H(a', b)\right) \;=\; \frac{1}{2|E|}\left(H(b', b) - H(a', b)\right). \end{aligned}$$

We can calculate μ_b directly as above, or use $\mu_b = 1 - \mu_a$ and (10). □

COROLLARY 14. *For a focal tree,* $T_{\text{forget}} = H(a', a)$. *For a bifocal tree,*

$$\begin{aligned} T_{\text{forget}} &= H(a', \mu) = H(b', \mu) \\ &= \frac{1}{2|E|}(H(a, b)H(b, a) + H(a, b)H(a', b) + H(b, a)H(b', a)). \end{aligned}$$

Proof. Since a' and b' are mixing pessimal, $T_{\text{forget}} = H(a', \mu) = H(b', \mu)$ and the first statement is obvious. If G is bifocal, then the following stopping rule is optimal

from a' to μ: walk until you hit b, then stop with probability μ_b, and walk to a with probability μ_a. Hence, $H(a',\mu) = H(a',b) + \mu_a H(b,a)$ and

$$T_{\text{forget}} = H(a',b) + \frac{H(b,a)}{2|E|}(H(b',b) - H(a',b)).$$

Equation (10) completes the proof. □

Proof of Theorem 8. We quickly narrow our search down to the foci of the tree. Recall that a stopping rule is optimal if and only if it has a halting state. Lovász and Winkler [10] show that every mixing pessimal node is a halting state for an optimal rule from μ to π. Hence on a tree, both a' and b' are halting states for an optimal rule from μ to π. Therefore, for any node i, the rule "follow an optimal rule from i to the forget distribution μ and then follow an optimal rule from μ to π" has either a' or b' as a halting state. This rule is optimal and $H(i,\pi) = H(i,\mu) + H(\mu,\pi)$. We may minimize $H(i,\mu)$ rather than $H(i,\pi)$, which is clearly minimized by a focus of the tree.

If G has a unique focus, then there is nothing to prove. Assume that G is bifocal with primary focus a and secondary focus b. Then

$$H(a,\mu) - H(b,\mu) = \mu_b H(a,b) - \mu_a H(b,a) = \frac{1}{2|E|}(H(a',b) - H(b',a))(H(a,b) + H(b,a)).$$

Thus $H(a,\mu) \geq H(b,\mu)$ if and only if $H(a',b) \geq H(b',a)$. □

5. Start-independent times. Start-independent stopping rules also identify central nodes: we now prove Proposition 9 and Theorem 10, which show that the target distributions achieving $T_{\text{si}}(\pi)$ and T_{si} are concentrated on a barycenter and the foci of the tree, respectively. The following lemma restricts our attention to singleton targets.

LEMMA 15. *Let σ and τ be distributions on the tree G. If τ has only one focus, then denote this node by u. Otherwise, let the foci u, v of τ satisfy $\sigma(V_{u:v})\pi(V_{u:v}) \geq \sigma(V_{v:u})\pi(V_{v:u})$. Then*

$$\sum_{k\in V} \sigma_k H(k,\tau) \geq \sum_{k\in V} \sigma_k H(k,u) = H(\sigma,u).$$

Proof. If u is the only focus for τ, then $\sum_{k\in V} \sigma_k H(k,\tau) = \sum_{k\in V} \sigma_k (H(k,u) + H(u,\tau)) \geq \sum_{k\in V} \sigma_k H(k,u)$. If τ has two foci with $\sigma(V_{u:v})\pi(V_{u:v}) \geq \sigma(V_{v:u})\pi(V_{v:u})$, then (9) implies that $\sigma(V_{u:v})H(u,v) \geq \sigma(V_{v:u})H(v,u)$. By Theorem 5,

$$\begin{aligned}
\sum_{k\in V} \sigma_k H(k,\tau) &= \sum_{k\in V_{u:v}} \sigma_k (H(k,u) + H(u,\tau)) + \sum_{k\in V_{v:u}} \sigma_k (H(k,v) + H(v,\tau)) \\
&= \sum_{k\in V_{u:v}} \sigma_k H(k,u) + \sum_{k\in V_{v:u}} \sigma_k H(k,v) + \sigma(V_{u:v})H(u,\tau) + \sigma(V_{v:u})H(v,\tau) \\
&= \sum_{k\in V} \sigma_k H(k,u) - \sigma(V_{v:u})H(v,u) + \sigma(V_{u:v})H(u,\tau) + \sigma(V_{v:u})H(v,\tau).
\end{aligned}$$

For any rule from u to τ, we must step from u to $V_{v:u}$ with probability $\tau(V_{v:u})$ before halting, hence $H(u,\tau) \geq \tau(V_{v:u})H(u,v)$. Likewise, $H(v,\tau) \geq \tau(V_{u:v})H(v,u)$. Therefore

$$\begin{aligned}
\sigma(V_{u:v})H(u,\tau) + \sigma(V_{v:u})H(v,\tau) &\geq \sigma(V_{u:v})\tau(V_{v:u})H(u,v) + \sigma(V_{v:u})\tau(V_{u:v})H(v,u) \\
&\geq \sigma(V_{v:u})H(v,u)
\end{aligned}$$

so $\sum_{k\in V}\sigma_k H(k,\tau) \geq \sum_{k\in V}\sigma_k H(k,u)$. ∎

Proof of Proposition 9. Taking $\sigma = \pi$ in Lemma 15, a singleton target achieves $T_{\mathrm{si}}(\pi)$. Proposition 1 shows that $T_{\mathrm{si}}(\pi) = \min_{i\in V}\sum_{k\in V}\pi_k H(k,i) = \min_{i\in V} H(\pi,i) = H(\pi,c)$. ∎

Obviously, $T_{\mathrm{si}}(\pi) \leq T_{\mathrm{reset}}$ and, in fact, $T_{\mathrm{si}}(\pi)$ can be arbitrarily small in comparison. Consider the tree consisting of a path of length $2k$ with k^4 leaves connected to the central node c. The focus, center, and barycenter of G are all located at c; therefore the forget distribution is concentrated on this central node. $T_{\mathrm{forget}} = k^2$ while $H(\pi,c)$ becomes arbitrarily close to $1/2$ for large k.

On the other hand, Theorem 10 states that the forget time and the start-independent time of a tree are identical. The theorem is clearly true for the path on two nodes, so we restrict our proof to trees on three or more nodes. We make the important observation that T_{si} need not be achieved by a unique pair of distributions. For example, consider a star graph with n leaves. Clearly, $T_{\mathrm{si}} \leq 1$ since we may always choose the central node c as our target. For any distribution σ concentrated on the leaf set such that $\sigma_i \leq (2n-1)/2n$ for every node i, we have $T_{\mathrm{si}} = \min_{j\in V} H(\sigma,j) = H(\sigma,c) = 1$.

We prove Theorem 10 by constructing a particular initial distribution ϕ concentrated on two leaves such that $\min_{j\in V} H(\phi,j) = T_{\mathrm{si}}$. Once we have identified such a ϕ, we show that we may choose the target node to be a focus of G.

LEMMA 16. *Given a distribution σ, the node u satisfies $H(\sigma,u) = \min_{j\in V} H(\sigma,j)$ if and only if for each neighbor v of u,*

$$\sigma(V_{v:u})H(v,u) \leq \sigma(V_{u:v})H(u,v) \tag{23}$$

or, equivalently,

$$\sigma(V_{v:u}) \leq \frac{H(u,v)}{2|E|} \text{ and } \sigma(V_{v:u}) \geq \frac{H(v,u)}{2|E|}. \tag{24}$$

We have equality if and only if $H(\sigma,u) = H(\sigma,v)$ so that v is also a best target for σ. Furthermore, at most one neighbor of u can satisfy (23) *with equality.*

Proof. For any neighbor v of u,

$$H(\sigma,v) = \sum_{k\in V_{u:v}}\sigma_k H(k,u) + \sigma(V_{u:v})H(u,v) + \sum_{k\in V_{v:u}}\sigma_k H(k,v).$$

We have $\sigma(V_{v:u})H(v,u) \leq \sigma(V_{u:v})H(u,v)$ if and only if

$$H(\sigma,v) \geq \sum_{k\in V_{v:u}}\sigma_k H(k,u) + \sigma(V_{v:u})H(v,u) + \sum_{k\in V_{v:u}}\sigma_k H(k,v) = H(\sigma,u).$$

Furthermore, equality holds in the first if and only if equality holds in the second. We find the equivalence of (23) and (24) by rewriting $\sigma(V_{v:u})H(v,u) \leq \sigma(V_{u:v})H(u,v) = (1-\sigma(V_{v:u}))H(u,v)$, solving for $\sigma(V_{v:u})$ and then using (10).

Now suppose that we have equality in (24) for two distinct u-neighbors v, w. Then

$$1 \geq \sigma(V_{v:u}) + \sigma(V_{w:u}) = \frac{1}{2|E|}\left(H(u,v)+H(u,w)\right) > \frac{1}{2|E|}\left(H(u,v)+H(v,u)\right) = 1,$$

a contradiction (where the inequality follows from (9)). ∎

We employ the following terminology for the remainder of this section. Let ϕ be a distribution, let $S_\phi = \{v|\phi_v > 0\} \subset V$, and let u be a node such that $H(\phi,u) =$

$\min_{j\in V} H(\phi, j) = T_{\mathrm{si}}$. Let $v_1, v_2, \ldots, v_{d(u)}$ be the neighbors of u, and let $w_i \in V_{v_i:u}$ be a leaf such that $H(w_i, u) = \max_{j\in V_{v_i:u}} H(j, u)$ for $1 \le i \le d(u)$.

Proof of Theorem 10. We first prove the result for stars $K_{1,k}$, $k \ge 0$. When $k = 0$ the result is trivial. Suppose that $G = K_{1,1}$ is the path on vertices u, v. Then $\max_\sigma \min_j H(\sigma, j) = \max_\sigma \min\{\sigma_u, \sigma_v\} = 1/2$. This value is achieved uniquely by taking $\sigma = (1/2, 1/2)$ and taking either node as the target.

Suppose that G is the star $K_{1,k}$, $k \ge 2$ with center c. When ϕ is divided evenly between two leaves u, v, then $\min_j H(\phi, j) = H(\phi, c) = 1$. We now show that $\min_j H(\sigma, j) \ge 1$ for any distribution σ. Note that $\min_j H(\sigma, j) \le H(\sigma, c) = 1 - \sigma_c \le 1$. So assume that a leaf w is the best target for σ. By Lemma 16, $\sigma(V_{c:w}) \le H(w, c)/2|E| = 1/2k$. In order to maximize $H(\sigma, w)$ we must set $\sigma_w = (2k-1)/2k$ and $\sigma_v = 1/2k$ for some leaf $v \ne w$. In this case $H(\sigma, w) = \frac{1}{2k} H(v, w) = 1$. Therefore $\max_\sigma \min_j H(\sigma, j) = 1$ for every star.

Now assume that G is a tree on four or more vertices and that G is not a star. Let $\sum_{i\in V} \phi_i H(i, u) = T_{\mathrm{si}}$.

Claim 1. The node u is not a leaf.

Assume that u is a leaf, and let v be its unique neighbor. Using (24), our best choice for the initial distribution is $\phi_u = 1/2|E|$ and $\phi_{u'} = 1 - \phi_u$. In this case, $H(\phi, u) = H(\phi, v)$ by Lemma 16, therefore v is also a minimizing target. We note that since G is not a star, $H(u', v) > 1$. Lemma 16 also guarantees that the remaining v-neighbors have a strict equality in (23). Therefore we can shift some weight from u to u' while still keeping v as the optimal target. Specifically, there exists some $\epsilon > 0$ such that the distribution ϕ' defined by $\phi'_u = \phi_u - \epsilon$, $\phi'_{u'} = \phi_{u'} + \epsilon$, and $\phi'_i = 0$ otherwise satisfies $\phi'(V_{w:v}) H(w, v) < \phi'(V_{v:w}) H(v, w)$ for every v-neighbor w. This ensures that v is the unique optimal target, while $H(\phi', v) > H(\phi, v) = H(\phi, u)$, a contradiction. Here the strict inequality follows from the fact that $H(u', v) > 1$.

Claim 2. S_ϕ intersects more than one component of $G\backslash\{u\}$.

Assume that S_ϕ intersects exactly one of $V_{v_1:u}, V_{v_2:u}, \ldots, V_{v_{d(u)}:u}$.

Case 1. $\phi_u = 0$. We may assume $S_\phi \subset V_{v_1:u}$. But $H(\phi, v_1) < H(\phi, u)$, contradicting $\min_{j\in V} H(\phi, j) = H(\phi, u)$.

Case 2. $\phi_u \ne 0$. We may assume $S_\phi \subset V_{v_1:u} \cup \{u\}$. By Lemma 16 we have $\phi(V_{v_1:u}) H(v_1, u) \le \phi(V_{u:v_1}) H(u, v_1) = \phi_u H(i, v_1)$. If we have equality here, then we may take v_1 as our target, proving the claim. If we have strict inequality, then there exists $\epsilon > 0$ such that the distribution ϕ' defined by $\phi'_u = \phi_u - \epsilon$, $\phi'_{v_1} = \phi_{v_1} + \epsilon$, and $\phi'_i = \phi_i$ otherwise satisfies $\phi'(V_{v_1:u}) H(v_1, u) < \phi'(V_{u:v_1}) H(u, v_1)$. Lemma 16 shows that $\min_{i\in V} H(\phi', i) = H(\phi', u)$, while $H(\phi', u) > H(\phi, u) = T_{\mathrm{si}}$, a contradiction.

Claim 3. ϕ may be chosen so that $S_\phi \subset \{w_1, w_2, \ldots, w_{d(u)}\}$.

Assume instead that $S_\phi \not\subset \{w_1, w_2, \ldots, w_{d(u)}\}$.

Case 1. $\phi_u = 0$. Let ϕ' be the distribution given by $\phi'_{w_i} = \phi(V_{v_i:u})$ for $1 \le i \le d(u)$ and zero elsewhere. Lemma 16 and $H(\phi, u) = \min_{i\in V} H(\phi, i)$ imply that $H(\phi', u) = \min_{i\in V} H(\phi', i)$ as well. Clearly, $H(\phi', u) \ge H(\phi, u)$, therefore we may use ϕ' in place of ϕ.

Case 2. $\phi_u \ne 0$. By an argument analogous to case 1, we may choose ϕ so that $S_\phi \subset \{u, w_1, w_2, \ldots, w_{d(u)}\}$. Suppose that $\phi(V_{u:v_i}) H(u, v_i) = \phi(V_{v_i:u}) H(v_i, u)$ for some i. We may take v_i as our target node in lieu of u by Lemma 16. Since v_i is an optimal target, Claim 1 ensures that v_i is not a leaf, therefore $\phi_{v_i} = 0$ and we have reduced ourselves to case 1.

If $\phi(V_{u:v_i}) H(u, v_i) > \phi(V_{v_i:u}) H(v_i, u)$ for $1 \le i \le d(u)$, then there exists $\epsilon > 0$ such that the distribution ϕ' given by $\phi'_u = \phi_u - \epsilon$, $\phi'_{w_1} = \phi_{w_1} + \epsilon$, and $\phi'_i =$

ϕ_i otherwise satisfies $\phi'(V_{v_i:u})H(v_i,u) < \phi'(V_{u:v_i})H(u,v_i)$ for $1 \le i \le d(u)$. By Lemma 16, $\min_{i\in V} H(\phi',i) = H(\phi',u) > H(\phi,u)$, a contradiction.

Claim 4. ϕ may be chosen to be concentrated on two leaves in $\{w_1, w_2, \dots, w_{d(u)}\}$.

Case 1. $\phi(V_{u:v_i})H(u,v_i) = \phi(V_{v_i:u})H(v_i,u)$ for some i. By Lemma 16, $H(\phi,u) = H(\phi,v_i)$. Notice that ϕ is supported in two components of $G\backslash\{v_i\}$, therefore using the proof of Claim 3 we may define a new distribution ϕ' concentrated on two leaves such that $\min_{k\in V} H(\phi',k) = H(\phi',v_i) \ge H(\phi,v_i) = H(\phi,u) = T_{\mathrm{si}}$.

Case 2. $\phi(V_{u:v_i})H(u,v_i) > \phi(V_{v_i:u})H(v_i,u)$ for all i. We show by induction that there exists a distribution ϕ' supported on two leaves such that $\min_{i\in V} H(\phi',i) \ge H(\phi,u)$. The base case $|S_\phi| = 2$ is trivial. Assume that if $|S_\phi| = k-1$, then there exists a ϕ' concentrated on two leaves satisfying $\min_{j\in V} \sum_{i\in V} \phi'_i H(i,j) = \sum_{i\in V} \phi'_i H(i,u) = \sum_{i\in V} \phi_i H(i,u) = T_{\mathrm{si}}$.

Considering $|S_\phi| = k \le d(u)$, order $S_\phi = \{w_1, w_2, \dots, w_k\}$ so that $H(w_1,u) \ge H(w_2,u) \ge \cdots \ge H(w_k,u)$. There exists $\epsilon > 0$ such that the distribution ϕ^* defined by $\phi^*_{w_1} = \phi_{w_1} + \epsilon$, $\phi^*_{w_k} = \phi_{w_k} - \epsilon$, and $\phi^*_i = \phi_i$ otherwise satisfies $\phi^*(V_{v_i:u})H(v_i,u) < \phi^*(V_{u:v_i})H(u,v_i)$ for all i. If $H(w_1,u) > H(w_k,u)$, then by Lemma 16, $\min_{i\in V} H(\phi^*,i) = H(\phi^*,u) > H(\phi,u) = T_{\mathrm{si}}$, a contradiction.

Otherwise, we have $H(w_i,u) = H(w_j,u)$ for $1 \le i,j \le k$. If there exists $0 < \epsilon < \phi_{w_k}$ such that $\phi^*(V_{u:v_k})H(u,v_k) = \phi^*(V_{v_k:u})H(v_k,u)$, then we have $H(\phi^*,v_k) = H(\phi^*,u) = H(\phi,u) = T_{\mathrm{si}}$. Hence we may take ϕ^* as our starting distribution and v_k as our target node. The support of ϕ^* is contained in two connected components of $V\backslash\{v_k\}$, and so the proof of Claim 3 shows that there exists a distribution ϕ' supported on two leaves such that $T_{\mathrm{si}} = \min_{i\in V} H(\phi',i)$. Finally, if $\phi^*(V_{u:v_k})H(u,v_k) > \phi^*(V_{v_k:u})H(v_k,u)$ for all $0 \le \epsilon \le \phi_{w_k}$, then by taking $\epsilon = \phi_{w_k}$ we have a distribution supported on $k-1$ leaves such that $\min_{i\in V} H(\phi^*,i) = T_{\mathrm{si}}$ and we are done by induction.

Claim 5. ϕ may be chosen so that S_ϕ is concentrated on two leaves w_1 and w_2 such that $H(w_1,u) \ge H(w_2,u)$ and the target node u is a focus of G. If the tree G is focal, then ϕ is concentrated on two u-pessimal leaves. If the tree G is bifocal, then w_1 is u-pessimal, v_1 is the other focus of G, and w_2 is v_1-pessimal. In this case, ϕ is given by $\phi_{w_1} = \pi(V_{u:v_1})$, $\phi_{w_2} = \pi(V_{v_1:u})$, and $H(\phi,v_1) = H(\phi,u) = T_{\mathrm{si}}$.

By Claim 4, we may assume that ϕ is concentrated on leaves $w_1 \in V_{v_1:u}$ and $w_2 \in V_{v_2:u}$ where $H(w_1,u) \ge H(w_2,u)$. In order to maximize the access time, the distribution ϕ must weight w_1 as much as possible while still keeping u as the best target node. By Lemma 16, we must have $\phi_{w_1}H(v_1,u) \le \phi_{w_2}H(u,v_1)$ and $\phi_{w_2}H(v_2,u) \le \phi_{w_1}H(u,v_2)$. By (9) this is equivalent to

$$\begin{cases} \phi_{w_1}\pi(V_{v_1:u}) \le \phi_{w_2}\pi(V_{u:v_1}), \\ \phi_{w_2}\pi(V_{v_2:u}) \le \phi_{w_1}\pi(V_{u:v_2}), \end{cases} \tag{25}$$

and the optimal choice is $\phi_{w_1} = \pi(V_{u:v_1})$ and $\phi_{w_2} = \pi(V_{v_1:u})$. Note that this choice results in

$$H(\phi,v_1) = H(\phi,u) \tag{26}$$

by Lemma 16.

The node w_1 is u-pessimal. Indeed by Claim 3, $H(w_1,u) = \max_{i\in V_{v_1:u}} H(i,u)$. Therefore if $H(w_1,u) < H(u',u)$, then a u-pessimal node u' must lie in one of $V_{v_i:u}$, $2 \le i \le d(u)$. Since $H(w_2,u) \le H(w_1,u)$, w_2 cannot be u-pessimal, therefore consider the distribution ϕ' given by $\phi'_{w_1} = \phi_{w_1}$, $\phi'_{u'} = \phi_{w_2}$, and $\phi'_i = 0$ otherwise.

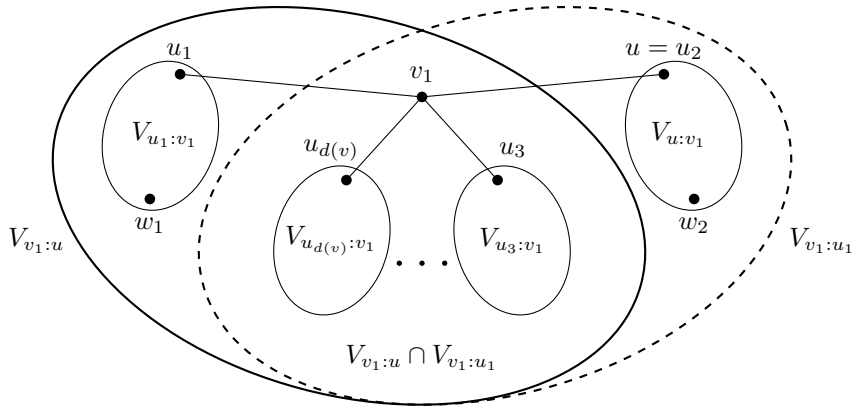


FIG. 2. *The tree decomposition in Claim* 5 *of the proof of Theorem* 10.

This distribution ϕ' satisfies the inequalities analogous to (25) so $\min_{i\in V} H(\phi', i) = H(\phi', u) > H(\phi, u) = T_{\text{si}}$, a contradiction. By a similar argument, w_2 must satisfy $H(w_2, u) = \max_{i\in V_{u:v_1}} H(i, u)$.

If u is the unique focus of G, then w_2 must also be u-pessimal. If u is a focus of a bifocal G, then v_1 must be the other focus of G and w_2 is v_1-pessimal. If u is not a focus, then the foci of G must be on the unique path between u and the u-pessimal node w_1. Hence w_1 is v_1-pessimal and $H(w_1, v_1) \geq H(w_2, v_1)$. By (26) we may take v_1 as our target node instead of u. If $H(w_1, v_1) = H(w_2, v_1)$, then v_1 must be the unique focus of G since w_1 is a v_1-pessimal node, and $T_{\text{si}} = H(\phi, v_1)$ as required.

Assume for the sake of contradiction that u is not a focus of G and $H(w_1, v_1) > H(w_2, v_1)$. Let $u_1, u_2, \ldots, u_{d(v_1)}$ be the neighbors of v_1, with $w_1 \in V_{u_1:v_1}$ and $w_2 \in V_{u_2:v_1}$. This ordering ensures that $u = u_2$. Consider the distribution ϕ' given by $\phi'_{w_1} = \pi(V_{v_1:u_1})$, $\phi'_{w_2} = \pi(V_{u_1:v_1})$, and $\phi'_i = 0$ otherwise. By Lemma 16, $H(\phi', v_1) = \min_{i\in V} H(\phi', i)$ and

$$\begin{aligned}
H(\phi', v_1) &= \pi(V_{v_1,u_1})H(w_1, v_1) + \pi(V_{u_1:v_1})H(w_2, v_1) \\
&= \left(\pi(V_{u:v_1}) + \pi(V_{v_1:u} \cap V_{v_1:u_1})\right) H(w_1, v_1) + \pi(V_{u_1:v_1})H(w_2, v_1) \\
&> \pi(V_{u:v_1})H(w_1, v_1) + \left(\pi(V_{v_1:u} \cap V_{v_1:u_1}) + \pi(V_{u_1:v_1})\right) H(w_2, v_1) \\
&= \pi(V_{u:v_1})H(w_1, v_1) + \pi(V_{v_1:u})H(w_2, v_1) \\
&= H(\phi, v_1) = H(\phi, u) = T_{\text{si}},
\end{aligned}$$

where the second and fourth equalities follow from the decomposition of the tree (as seen in Figure 2), and the inequality is due to $H(w_1, v_1) > H(w_2, v_2)$. The resulting inequality $H(\phi', v_1) > T_{\text{si}}$ is a contradiction, therefore u must be a focus of G.

Completion of proof. If G has a single focus a, then by Claim 5 we may take ϕ to be concentrated on two a-pessimal leaves in different components of $G\backslash\{a\}$ and $T_{\text{si}} = H(\phi, a) = H(a', a) = T_{\text{forget}}$ by Corollary 14. If G is bifocal with foci a and b, then by Claim 5 we may take ϕ to be concentrated on an a-pessimal node a' and a b-pessimal node b'. Also, $T_{\text{si}} = H(\phi, a) = H(\phi, b)$ by (26). Finally,

$$T_{\text{si}} = (\mu_a + \mu_b)T_{\text{si}} = \mu_a H(\phi, a) + \mu_b H(\phi, b)$$

$$= \phi_{a'}(\mu_a H(a', a) + \mu_b H(a', b)) + \phi_{b'}(\mu_a H(b', a) + \mu_b H(b', b))$$

$$= \phi_{a'} H(a', \mu) + \phi_{b'} H(b', \mu) = (\phi_{a'} + \phi_{b'}) T_{\text{forget}} = T_{\text{forget}}$$

by Proposition 7. □

Acknowledgments. The author would like to thank László Lovász for many insightful conversations, Peter Winkler for his helpful comments, and the anonymous referees for their suggested improvements to the exposition.

REFERENCES

[1] D. Aldous, *Some inequalities for reversible Markov chains*, J. London Math. Soc. (2), 25 (1982), pp. 564–576.
[2] D. Aldous, L. Lovász, and P. Winkler, *Mixing times for uniformly ergodic Markov chains*, Stochastic Process. Appl., 71 (1997), pp. 165–185.
[3] A. Beveridge and L. Lovász, *Random walks and the regeneration time*, J. Graph Theory, 29 (1998), pp. 57–62.
[4] G. Brightwell and P. Winkler, *Extremal cover times for random walks on trees*, J. Graph Theory, 14 (1990), pp. 547–554.
[5] D. Coppersmith, P. Tetali, and P. Winkler, *Collisions among random walks on a graph*, SIAM J. Discrete Math., 6 (1993), pp. 363–374.
[6] P. Diaconis and J. A. Fill, *Strong stationary times via a new form of duality*, Ann. Probab., 18 (1990), pp. 1483–1522.
[7] I. Dumitriu, P. Tetali, and P. Winkler, *On playing golf with two balls*, SIAM J. Discrete Math., 16 (2003), pp. 604–615.
[8] L. Lovász, *Random walks on graphs: A survey*, in Combinatorics, Paul Erdős is Eighty, Vol. II, D. Miklós, V. T. Sós, and T. Szőnyi, eds., J. Bolyai Math. Soc., Budapest, 1996, pp. 353–397.
[9] L. Lovász and P. Winkler, *Efficient stopping rules for Markov chains*, in Proceedings of the 27th ACM Symposium on the Theory of Computing, ACM, New York, 1995, pp. 76–82.
[10] L. Lovász and P. Winkler, *Reversal of Markov chains and the forget time*, Combin., Probab. Comput., 7 (1998), pp. 189–204.
[11] J. W. Moon, *Random walks on random trees*, J. Austral. Math. Soc., 15 (1973), pp. 42–53.
[12] L. H. Pearce, *Random walks on trees*, Discrete Math., 30 (1980), pp. 269–276.
[13] J. W. Pitman, *Occupation measures for Markov chains*, Advances in Appl. Probability, 9 (1977), pp. 69–86.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 319–332

A COMBINATORIAL CHARACTERIZATION OF COXETER GROUPS*

MARIO MARIETTI†

Abstract. In this paper we give a purely combinatorial characterization of Coxeter groups in terms of their partial order structure under Bruhat order. The result is based on the recently introduced concept of special matching and is achieved by proving an analogue of Tits' word theorem. As a consequence of the proof of our main result, we obtain a result about shellability.

Key words. Coxeter groups, Bruhat order, special matchings

AMS subject classifications. 20F55, 05E15

DOI. 10.1137/070695034

1. Introduction. Coxeter group theory derives much of its appeal from its interactions with several areas of mathematics such as algebra, combinatorics, and geometry (see, e.g., [1], [7], [14], [15]). In Coxeter group theory, a crucial role is played by Bruhat order. It arises not only in the Bruhat decomposition (this motivates the terminology although it would be more appropriate to call it Chevalley order) but also in many other contexts such as in connection with inclusions among Schubert varieties, with the Verma modules of a complex semisimple Lie algebra, and in Kazhdan–Lusztig theory.

The problem of characterizing Coxeter groups among the groups generated by involutions and the problem of characterizing the Bruhat order among the partial orders on a fixed Coxeter group have been studied intensively and solved (see, e.g., [1], [7], [10], [11], [13], [15], [18]). In this paper we solve the problem of giving a characterization of Coxeter groups partially ordered by Bruhat order among all possible partially ordered sets (or posets for short). In other words, we give a necessary and sufficient condition for an abstract poset to be isomorphic to a Coxeter group partially ordered by Bruhat order (Theorem 4.2). This result is proved by studying the combinatorics of words in the alphabet of special matchings and, in particular, giving a combinatorial version of Tits' word theorem. As a consequence of our main result, we describe the combinatorial relation between (strong) Bruhat order and weak Bruhat order, and we can prove that certain labelings arising from words of special matchings are CL-labelings (chain lexicographical labelings). Hence it follows that certain posets studied in [17] are shellable.

The paper is organized as follows. In section 2, we recall some basic definitions and results that are needed in what follows. In section 3, we study the combinatorics of words in the alphabet of special matchings on a class of posets called zircons. In particular, we can introduce (reduced) expressions, exchange condition and subword property also in the context of zircons. In section 4, using the results in section 3, we prove a combinatorial version of Tits' word theorem for words of special matchings, and we give the main result of the paper, namely, a characterization of Coxeter groups partially ordered by Bruhat order. Furthermore, we characterize the Bruhat order

*Received by the editors June 21, 2007; accepted for publication (in revised form) September 9, 2008; published electronically January 7, 2009.
http://www.siam.org/journals/sidma/23-1/69503.html
†Dipartimento di Matematica, Sapienza Università di Roma, Piazzale Aldo Moro 5, 00185 Roma, Italy (marietti@mat.uniroma1.it).

among all partial orders on W as the unique order for which the multiplication by a fixed generator gives rise to a special matching. As a corollary, we describe a combinatorial way to obtain the weak Bruhat order from the Bruhat order and vice versa. In section 5, we give a result about shellability of zircons.

2. Notation and background. This section reviews the background material on posets, Coxeter systems, and special matchings that is needed in the rest of this paper. We refer the reader to [1], [15], and [16] for a more detailed treatment. We write ":=" if we are defining the left-hand side by the right-hand side. We let $\mathbb{N} := \{0, 1, 2, 3, \dots\}$, and for $a, b \in \mathbb{N}$ we let $[a, b] := \{a, a+1, a+2, \dots, b\}$ and $[a] := \{1, 2, \dots, a\}$. The cardinality of a set A will be denoted by $|A|$.

Let P be a poset. An element $x \in P$ is *maximal* (respectively, *minimal*) if there is no element $y \in P \setminus \{x\}$ such that $x \leq y$ (respectively, $y \leq x$). We say that P has a *bottom element* $\hat{0}$ if there exists an element $\hat{0} \in P$ satisfying $\hat{0} \leq x$ for all $x \in P$. Similarly, P has a *top element* $\hat{1}$ if there exists an element $\hat{1} \in P$ satisfying $x \leq \hat{1}$ for all $x \in P$. If $x \leq y$, we define the *(closed) interval* $[x, y] = \{z \in P : x \leq z \leq y\}$ and the *open interval* $(x, y) = \{z \in P : x < z < y\}$. If every interval of P is finite, then P is called a *locally finite poset*. We say that x is covered by y if $x < y$ and $[x, y] = \{x, y\}$, and we write $x \lhd y$ or $y \rhd x$. If P has a $\hat{0}$, then an element $x \in P$ is an *atom* of P if $\hat{0} \lhd x$. A *chain* of P is a totally ordered subset of P. A chain c with top element y and bottom element x is *saturated* if it is a maximal chain of the interval $[x, y]$. A poset P is *ranked* if there exists a (rank) function $\rho : P \to \mathbb{N}$ such that $\rho(y) = \rho(x) + 1$ whenever $x \lhd y$. A poset P is *pure* of length $\ell(P) = n$ if all maximal chains are of the same length n. A poset P with bottom element $\hat{0}$ is *graded* if every interval $[\hat{0}, x]$ is pure. A poset P is a Boolean algebra if it is isomorphic to the poset of all subsets of a certain set S, partially ordered by inclusion.

A standard way of depicting a poset P is by its *Hasse diagram*. This is the digraph with P as node set and having an upward-directed edge from x to y if and only if $x \lhd y$. We say that P is *connected* if its Hasse diagram is connected as a graph. A morphism of posets is a map $\phi : P \to Q$ from the poset P to the poset Q which is *order-preserving*, namely, such that $x \leq y$ in P implies $\phi(x) \leq \phi(y)$ in Q for all $x, y \in P$. Two posets P and Q are *isomorphic* if there exists an order-preserving bijection $\phi : P \to Q$ whose inverse is also order-preserving. In this case ϕ is an isomorphism of posets.

We follow [1] for undefined Coxeter groups notation and terminology. Given a Coxeter system (W, S), we denote by $(m(s, t))_{s,t \in S}$ the Coxeter matrix of W. Given $w \in W$, we denote by $l(w)$ the length of w, we call any product of $l(w)$ elements of S which represents w a *reduced expression for* w, and we let

$$D_R(w) := \{s \in S : \ l(ws) < l(w)\} = D_L(w^{-1}),$$
$$D_L(w) := \{s \in S : \ l(sw) < l(w)\} = D_R(w^{-1}).$$

We call $D_R(w)$ and $D_L(w)$, respectively, the *right* and the *left descent set of* w. We denote by e the identity of W, and we let $T := \{wsw^{-1} : w \in W, \ s \in S\}$ be the set of reflections of W.

The following property, known as the exchange condition, characterizes the Coxeter groups among the groups generated by involutions.

THEOREM 2.1. *Let $w \in W$, and let $s_1 s_2 \dots s_r$ be a reduced expression for w. Let $s \in S$ be such that $l(ws) < l(w)$. Then there exists a unique $i \in [r]$ such that $ws = s_1 s_2 \dots \widehat{s_i} \dots s_r$ (where $\widehat{s_i}$ means that s_i has been deleted). Furthermore, the positive integer i is such that $s_{i+1} s_{i+2} \dots s_r s$ is reduced, while $s_i s_{i+1} \dots s_r s$ is not.*

We now recall a result due to Tits [19] that is needed in what follows. Given $s,t \in S$ such that $m(s,t) < \infty$, let $\alpha_{s,t} = ststst\ldots$ with exactly $m(s,t)$ letters. Two expressions are said to be linked by a braid move (respectively, a nil move) if it is possible to obtain the first from the second by changing a factor $\alpha_{s,t}$ to a factor $\alpha_{t,s}$ (respectively, by deleting a factor ss).

THEOREM 2.2 (Tits' word theorem). *Let $w \in W$. Then any expression for w (not necessarily reduced) is linked to any reduced expression for w by a finite sequence of braid and nil moves.*

The Coxeter group W is partially ordered by *(strong) Bruhat order*. Given $u, v \in W$, $u \leq v$ if and only if there exist $r \in \mathbb{N}$ and $t_1, \ldots, t_r \in T$ such that $t_r \ldots t_1\, u = v$ and $l(t_i \ldots t_1 u) > l(t_{i-1} \ldots t_1 u)$ for $i = 1, \ldots, r$. It is well known that W, partially ordered by Bruhat order, is a graded poset having l as its rank function. There is a well-known characterization of Bruhat order on a Coxeter group in terms of reduced expressions (usually referred to as the *subword property*). By a *subword* of a word $s_1 s_2 \ldots s_q$ we mean a word of the form $s_{i_1} s_{i_2} \ldots s_{i_k}$, where $1 \leq i_1 < \cdots < i_k \leq q$.

THEOREM 2.3. *Let $u, v \in W$. Then the following are equivalent:*

1. *$u \leq v$;*
2. *every reduced expression for v has a subword that is a reduced expression for u;*
3. *there exists a reduced expression for v which has a subword that is a reduced expression for u.*

Another partial order structure on the Coxeter group W is given by the *(weak) Bruhat order* that we denote by $\leq_w$. Given $u, v \in W$, $u \leq_w v$ if and only if there exist $r \in \mathbb{N}$ and $s_1, \ldots, s_r \in S$ such that $s_r \ldots s_1\, u = v$ and $l(s_i \ldots s_1 u) > l(s_{i-1} \ldots s_1 u)$ for $i = 1, \ldots, r$. A characterization of the poset $(W, \leq_w)$ has been given by Eriksson [12].

THEOREM 2.4. *For any Coxeter group (W,S) with Coxeter matrix $(m(s,t))_{s,t \in S}$, there exists a unique poset P (up to isomorphism) such that*

1. *P has a bottom element $\hat{0}$;*
2. *P has $|S|$ atoms;*
3. *P admits a labeling of the edges of its Hasse diagram with labels in S satisfying the following:*
   - *no two edges incident to the same element of P have the same label;*
   - *if there are two edges going upwards from an element $p \in P$ with labels s and t, then they are the first edges of two upward-going paths from p of length $m(s,t)$ labeled alternatingly s and t. If $m(s,t) < \infty$, then these paths end in the same element; while if $m(s,t) = \infty$, the paths go on forever.*

Such a poset P is isomorphic to $(W, \leq_w)$.

Recall that a matching of a graph $G = (V, E)$ is an involution $M : V \to V$ such that $\{M(v), v\} \in E$ for all $v \in V$. Let P be a poset. A matching M of the Hasse diagram of P is a *special matching of P* if

$$u \lhd v \Longrightarrow M(u) \leq M(v)$$

for all $u, v \in P$ such that $M(u) \neq v$.

For the reader's convenience, we collect the following two results. The first one appears in [9], while the second one follows easily by Lemma 4.2 of [8].

LEMMA 2.5. *Let P be a locally finite ranked poset, M be a special matching of P, and $u, v \in P$, $u \leq v$, be such that $M(u) \rhd u$ and $M(v) \lhd v$. Then M restricts to a special matching of $[u, v]$.*

LEMMA 2.6 (lifting lemma for special matchings). *Let M be a special matching of a locally finite ranked poset P, and let $u, v \in P$, $u \leq v$. Then*

1. *if $M(v) \lhd v$ and $M(u) \lhd u$, then $M(u) \leq M(v)$;*
2. *if $M(v) \rhd v$ and $M(u) \rhd u$, then $M(u) \leq M(v)$;*
3. *if $M(v) \lhd v$ and $M(u) \rhd u$, then $M(u) \leq v$ and $u \leq M(v)$.*

Given an element $w \in P$, we say that M is a special matching of w if M is a special matching of the subposet $\{x \in P : x \leq w\}$. We denote the set of all special matchings of w by SM_w. Recall from [17] the following definition.

DEFINITION 2.7. *We say that a locally finite ranked poset Z is a zircon if SM_w is nonempty for all $w \in Z$, w not minimal.*

Note that the set SM_Z of all special matchings of the entire zircon Z may happen to be empty.

The following assertions are proved in [17].

THEOREM 2.8. *Any zircon is a disjoint union of graded posets (its connected components). Any connected zircon is a Eulerian poset. Any interval of length 3 in a zircon is a k-crown. Any interval of length 2 is a square (namely, it has 4 elements).*

All Coxeter groups partially ordered by Bruhat order are connected zircons. In fact, let (W, S) be any Coxeter system. Then W is a locally finite ranked poset with the length function as rank function. Fix $w \in W \setminus \{e\}$ and $s \in D_R(w)$. Then the involution $\rho_s : [e, w] \to [e, w]$ defined by $\rho_s(u) = us$ for all $u \in [e, w]$ is a special matching of w. Similarly, if $s \in D_L(w)$, the involution $\lambda_s : [e, w] \to [e, w]$ defined by $\lambda_s(u) = su$ for all $u \in [e, w]$ is a special matching of w.

3. Words and special matchings in zircons. In this section we show that the special matchings of a zircon play the role that Coxeter generators play in Coxeter group theory. Throughout this section, we let Z be a connected zircon with rank function ρ (by Theorem 2.8 no generality is lost), and we let $SM(Z) = \cup_{z \in Z} SM_z$.

DEFINITION 3.1. *For any $z \in Z$, we say that a word $M_1 \dots M_\rho$ in the alphabet $SM(Z)$ is an* expression of special matchings for z *if $M_i \dots M_\rho(z)$ is defined for all $i \in [\rho]$ and $M_1 \dots M_\rho(z) = \hat{0}$ (or, equivalently, if $M_i \dots M_1(\hat{0})$ is defined for all $i \in [\rho]$ and $M_\rho \dots M_1(\hat{0}) = z$). Furthermore, we say that the expression $M_1 \dots M_\rho$ for z is* reduced *if $\rho = \rho(z)$.*

Note that if $M_1 \dots M_\rho$ is a reduced expression for z, then, for all $i \in [\rho]$,

$$M_i M_{i+1} \dots M_\rho(z) \lhd M_{i+1} \dots M_\rho(z),$$

and M_i restricts to a special matching of $M_{i+1} \dots M_\rho(z)$ by Lemma 2.5. Then we can associate to any reduced expression for z a regular sequence of special matchings for z in the sense of section 9 of [9]. It is sometimes useful to consider the chain associated to a reduced expression $M_1 \dots M_\rho$ for an element v. This is the chain $(v_0, \dots, v_\rho)$ where $v_i := M_{i+1} \dots M_\rho(v) = M_i \dots M_1(\hat{0})$ for $i = 0, \dots, \rho$. For example, consider the zircon in Figure 1 and its special matchings M, N, and O. The expression $MONM$ is a reduced expression of special matchings for v_4 with $(v_0, v_1, v_2, v_3, v_4)$ as the associated chain.

Remark. Let w be an element in a Coxeter group W, and let $s_1 \dots s_\ell$ be an expression (respectively, a reduced expression) for w in the Coxeter group terminology. Then $\lambda_{s_\ell} \dots \lambda_{s_1}$ and $\rho_{s_1} \dots \rho_{s_\ell}$ are both expressions (respectively, both reduced expressions) of special matchings for w (see the remark at the end of section 2 for the notation). Thus, the concept of (reduced) expressions of special matchings in zircons is a generalization of that of (reduced) expressions in Coxeter group theory.

The proof of the following result is similar to the proof of the corresponding result for regular sequences in Coxeter groups (Lemma 9.2 of [9]) and is therefore omitted.

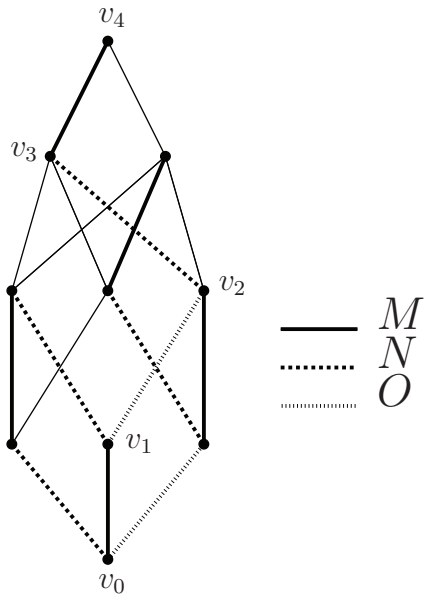


FIG. 1. *$MONM$ is a reduced expression of special matchings.*

By a *subword* of a word $M_1M_2 \dots M_r$ we mean a word of the form $M_{i_1}M_{i_2} \dots M_{i_k}$, where $1 \le i_1 < \cdots < i_k \le r$.

LEMMA 3.2. *Let $z \in Z$, and let $M_1 \dots M_\rho$ be a reduced expression of special matchings for z. Then the composition $M_{i_k} \dots M_{i_1}(\hat{0})$ is defined for any subword $M_{i_1} \dots M_{i_k}$ of $M_1 \dots M_\rho$ and $M_{i_k} \dots M_{i_1}(\hat{0}) \le z$.*

PROPOSITION 3.3. *Let $u, v \in Z$, $u \lhd v$, and let $M_1 \dots M_\rho$ be a reduced expression for v. Then there exists a unique $i \in [\rho]$ such that $u = M_\rho M_{\rho-1} \dots M_i \dots M_{\rho-1}M_\rho(v)$. Furthermore,*

$$i = \max\{k \in [\rho] : M_k \dots M_{\rho-1}M_\rho(u) \ge M_{k+1} \dots M_{\rho-1}M_\rho(u)\}.$$

Proof. To prove the existence part, we prove that the assertion holds for $i = \max\{k \in [\rho] : M_k \dots M_{\rho-1}M_\rho(u) \ge M_{k+1} \dots M_{\rho-1}M_\rho(u)\}$. We proceed by induction on the rank ρ of v, the assertion being clear if $\rho = 1$.

So suppose $\rho > 1$. If $M_\rho(v) = u$, then $i = \rho$, and we are done. Otherwise, by the definition of a special matching, $M_\rho(u) \lhd u$ and $M_\rho(u) \lhd M_\rho(v)$. As $M_1 \dots M_{\rho-1}$ is a reduced expression for $M_\rho(v)$, by the induction hypothesis, we have that $M_\rho(v) = M_{\rho-1}M_{\rho-2} \dots M_{i'} \dots M_{\rho-2}M_{\rho-2}(M_\rho(u))$, where

$$i' = \max\{k \in [\rho-1] : M_k \dots M_{\rho-1}(M_\rho(u)) \ge M_{k+1} \dots M_{\rho-1}(M_\rho(u))\}.$$

Clearly $i' = i$ and then $v = M_\rho M_{\rho-1} \dots M_i \dots M_{\rho-1}M_\rho(u)$ since the special matchings are involutions. So we are done.

Let us prove the uniqueness part. By contradiction, suppose that, for some $k \ne i$, $u = M_\rho M_{\rho-1} \dots M_k \dots M_{\rho-1}M_\rho(v)$. Then $k < i$ because $M_\rho M_{\rho-1} \dots M_k \dots M_{\rho-1}$ $M_\rho(u)$ has rank $< \rho(v)$ for all $k > i$. Since we have proved that $u = M_\rho M_{\rho-1} \dots M_i \dots$ $M_{\rho-1}M_\rho(v)$, it follows that

$$M_{i+1} \dots M_{\rho-1}M_\rho(v) = M_{i-1} \dots M_{k+1}M_kM_{k+1} \dots M_{i-1}M_iM_{i+1} \dots M_{\rho-1}M_\rho(v).$$

Hence we have

$$M_{k+1} \dots M_{i-1} M_{i+1} \dots M_{\rho-1} M_\rho(v) = M_k \dots M_{i-1} M_i M_{i+1} \dots M_{\rho-1} M_\rho(v).$$

This is impossible since either the left-hand side is not defined or its rank is greater than the rank of the right-hand side. ∎

Remark. Let (W, S) be a Coxeter system, $v \in W$, and let $s_1 \dots s_r$ be a reduced expression for v in the Coxeter group terminology. It is a well-known fact that $\{t \in T : l(tv) < l(v)\} = \{s_1 s_2 \dots s_i \dots s_2 s_1 : i \in [r]\}$. Hence the Bruhat order on W is actually defined as the transitive closure of the covering relations of the form in Proposition 3.3 in the case where the expression $M_1 \dots M_\rho$ comes from a reduced expression in the Coxeter group terminology.

The following results are the generalizations of Theorems 2.1 and 2.3 to arbitrary zircons.

THEOREM 3.4 (exchange condition for zircons). *Let $v \in Z$, and let $M_1 \dots M_\rho$ be a reduced expression for v. Then, for all $u \lhd v$, there exists a unique $i \in [\rho]$ such that $M_1 \dots \widehat{M_i} \dots M_\rho$ (M_i deleted) is a reduced expression for u. Furthermore,*

$$i = \max\{k \in [\rho] : M_k \dots M_{\rho-1} M_\rho(u) \geq M_{k+1} \dots M_{\rho-1} M_\rho(u)\}.$$

Proof. By Proposition 3.3, $u = M_\rho M_{\rho-1} \dots M_i \dots M_{\rho-1} M_\rho(v)$, and hence we have $M_1 \dots \widehat{M_i} \dots M_\rho(u) = M_1 \dots M_\rho(v) = \hat{0}$ since the special matchings are involutions.

The uniqueness part also follows by Proposition 3.3. ∎

THEOREM 3.5 (subword property for zircons). *Let $u, v \in Z$. Then the following are equivalent:*

1. *$u \leq v$;*
2. *every reduced expression for v has a subword that is a reduced expression for u;*
3. *there exists a reduced expression for v which has a subword that is a reduced expression for u.*

Proof. Let us first show that item 1 implies item 2. We proceed by induction on $\rho := \rho(v)$, the statement being trivial for $\rho = 1$. So assume that $\rho > 1$, and fix a reduced expression $M_1 \dots M_\rho$ for v. Clearly $M_1 \dots M_{\rho-1}$ is a reduced expression for $M_\rho(v)$. If $M_\rho(u) \lhd u$, then, by Lemma 2.6, $M_\rho(u) \leq M_\rho(v)$. So by induction there exist $1 \leq i_1 < \dots < i_k \leq \rho - 1$ such that $M_{i_1} \dots M_{i_k}$ is a reduced expression for $M_\rho(u)$, and hence $M_{i_1} \dots M_{i_k} M_\rho$ is a reduced expression for u. If $M_\rho(u) \rhd u$, then, by Lemma 2.6, $u \leq M_\rho(v)$, and we again conclude by induction.

Clearly item 2 implies item 3.

To prove that item 3 implies item 1 we proceed by induction on $\rho := \rho(v)$, the claim being clear if $\rho = 1$. Let $M_1 \dots M_\rho$ be a reduced expression for v, and suppose that $M_{i_1} \dots M_{i_k}$ is a reduced expression for u. In particular, $u = M_{i_k} \dots M_{i_1}(\hat{0})$. If $i_k \neq \rho$, then $M_{i_1} \dots M_{i_k}$ is a subword of $M_1 \dots M_{\rho-1}$, which is a reduced expression for $M_\rho(v)$. Hence, by our induction hypothesis, $u \leq M_\rho(v) < v$, and we are done. Suppose now that $i_k = \rho$. Clearly $M_{i_1} \dots M_{i_{k-1}}$ is both a subword of $M_1 \dots M_{\rho-1}$ and a reduced expression for $M_\rho(u)$. Then, by our induction hypothesis, $M_\rho(u) \leq M_\rho(v)$, and hence, by Lemma 2.6, we get the assertion. ∎

The next result shows that, given $u \leq v \in Z$ and a reduced expression for v, there are two canonical ways of choosing a subword of this reduced expression which is a reduced expression for u.

LEMMA 3.6. *Let $v \in Z$, $\mathcal{M} = M_1 \dots M_\rho$ be a reduced expression for v, and $(v_0, \dots, v_\rho)$ be the chain associated to $\mathcal{M}$. Then, for all $u \le v$, u not minimal, we have*

1. *$\{j \in [\rho] : M_j(u)$ is defined and $M_j(u) \lhd u\} \neq \emptyset$;*
2. *$M_k(u) \lhd u$, where $k := \min\{j \in [\rho] : v_j \ge u\}$.*

Proof of item 1. We proceed by induction on ρ, the claim being clear if $\rho = 1$. Assume $\rho > 1$, and consider the special matching M_ρ of v. Clearly $M_\rho(u)$ is defined since $u \le v$. If $M_\rho(u) \lhd u$, then we are done. Otherwise, by Lemma 2.6, $u \le M_\rho(v)$, and we can conclude by induction since $M_1 \dots M_{\rho-1}$ is a reduced expression for $M_\rho(v)$. □

Proof of item 2. Clearly $M_k(u)$ is defined since $u \le v_k$. By contradiction, suppose that $M_k(u) \rhd u$. Then, by Lemma 2.6, $u \le M_k(v_k) = v_{k-1}$. This contradicts the minimality of k. □

After Lemma 3.6, given $v \in Z$ and a reduced expression $\mathcal{M} = M_1 \dots M_\rho$ for it, we let

$$\max_u = \max\{j \in [\rho] : M_j(u) \text{ is defined and } M_j(u) \lhd u\},$$
$$\min_u = \min\{j \in [\rho] : v_j \ge u\},$$

for all $u \le v$, u not minimal. We say that $M_{\max_u}$ and $M_{\min_u}$ are, respectively, the maximal special matching and the minimal special matching of u according to $\mathcal{M}$. The following result is an immediate consequence of Lemma 3.6.

COROLLARY 3.7. *Let $v \in Z$, $\rho := \rho(v)$, and $M_1 \dots M_\rho$ be a reduced expression for v. Then there are two canonical injective maps $i_{\max}, i_{\min} : [\hat{0}, v] \to \{$subwords of $M_1 \dots M_\rho\}$ sending u to its unique reduced expression made of, respectively, maximal and minimal special matchings. In particular, $|[\hat{0}, v]| \le 2^\rho$ and $\{z \in [\hat{0}, v] : \rho(z) = k\} \le \binom{\rho}{k}$.*

Remark. The inequalities in Corollary 3.7 are sharp (see the Boolean algebra).

4. Main results. In this section we give a characterization for an abstract poset to be isomorphic to a Coxeter group partially ordered by Bruhat order. The class of zircons is much larger than the class of posets isomorphic to a Coxeter group. The poset in Figure 1 is the simplest zircon with a top element which is not isomorphic to a Bruhat interval in any Coxeter group. More generally, by Theorem 3.2 of [9], a poset P which does not avoid $K_{3,2}$ cannot be a Bruhat interval (P avoids $K_{3,2}$ if there are no elements $a_1, a_2, a_3, b_1, b_2 \in P$, all distinct, such that either $a_i \lhd b_j$ for all $i \in [3]$, $j \in [2]$, or $a_i \rhd b_j$ for all $i \in [3]$, $j \in [2]$). The zircon in Figure 2 has the distinguishing characteristic of admitting only one special matching of the entire poset. This cannot happen for a nontrivial interval $[e, w]$ in a Coxeter group W since any element $w \in W$ of length > 1 has at least one left descent and one right descent which give rise to two different special matchings.

The main step in proving the main result of this work is the following theorem, whose proof uses the results in the previous section and which gives a combinatorial version of Tits' word theorem for a certain class of zircons. Given two special matchings M, N of a poset P, we denote by $\langle M, N\rangle$ the subgroup of the symmetric group on P generated by M and N.

THEOREM 4.1. *Let P be a graded poset such that there exists a set $\mathcal{R}$ of special matchings of P satisfying the following:*

1. *for all $z \in P \setminus \{\hat{0}\}$, there exists $R \in \mathcal{R}$ such that $R(z) \lhd z$;*
2. *for all $R, R' \in \mathcal{R}$, all orbits under the action of $\langle R, R'\rangle$ have the same cardinality (possibly ∞).*

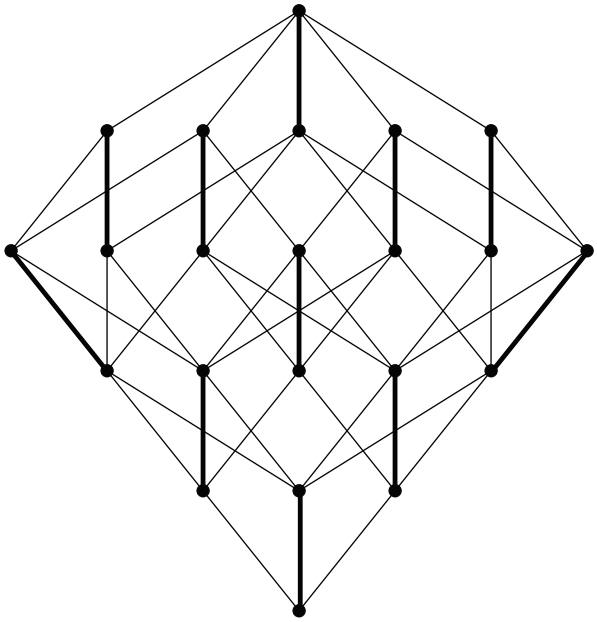


FIG. 2. *A zircon with only one special matching.*

Then P is a zircon, and given any expression of special matchings in $\mathcal{R}$ and any reduced expression of special matchings in $\mathcal{R}$, they are expressions for the same element if and only if one can reach the second from the first by a finite sequence of the following moves:

- *braid move: replace $\underbrace{\cdots RR'R}_{m(R,R')}$ with $\underbrace{\cdots R'RR'}_{m(R,R')}$ for some $R, R' \in \mathcal{R}$, where $2m(R, R')$ is the common cardinality of all orbits under the action of $\langle R, R' \rangle$;*
- *nil move: replace RR with the empty word for some $R \in \mathcal{R}$.*

Proof. Let us show that P is a zircon. By property 1 and Lemma 2.5, SM_w is nonempty for all $w \in P$. We still have to prove that P is locally finite. Let us proceed by contradiction. So let $w \in P$ be minimal such that $[\hat{0}, w]$ has infinite cardinality, and let ρ be the rank of w. Hence

$$\{j \in [\rho] : \text{there are an infinite number of elements of rank } j \text{ in } [\hat{0}, w]\} \neq \emptyset.$$

This set contains $\rho - 1$ since otherwise there would be an element of rank $< \rho$ covering an infinite number of elements, which contradicts the minimality of w. By property 1 there exists $R \in \mathcal{R}$ such that $R(w) \lhd w$. Then, by the definition of a special matching, $R(w)$ covers $R(v)$ for all $v \lhd w$, $v \neq R(w)$. But these are infinitely many, and we again conclude by contradiction.

Let us prove the second statement. By property 2 and by the fact that special matchings are involutions, the "if" part is clear. To prove the "only if" part, we first assume that the two expressions are both reduced expressions for $z \in Z$ and we proceed by induction on the rank $\rho(z)$ of z. So let $R_1 R_2 \dots R_{\rho(z)}$ and $T_1 T_2 \dots T_{\rho(z)}$ be two reduced expressions for z, and let $R := R_{\rho(z)}$, $T := T_{\rho(z)}$. If $R = T$, the assertion follows by induction considering the two reduced expressions $R_1 R_2 \dots R_{\rho(z)-1}$ and $T_1 T_2 \dots T_{\rho(z)-1}$. If $R \neq T$, z is the top element of its orbit under the action of $\langle R, T \rangle$. Let x be the bottom element of this orbit, and let $R_{c_1} R_{c_2} \dots R_{c_{\rho(x)}}$ be a

reduced expression for x of special matchings in $\mathcal{R}$. Then $R_{c_1}R_{c_2}\dots R_{c_{\rho(x)}}\underbrace{\cdots RTR}_{m(R,T)}$ is a reduced expression for z of special matchings in $\mathcal{R}$. So, by the induction hypothesis, we have that $R_1R_2\dots R_{\rho(z)}$ is linked by braid moves to $R_{c_1}R_{c_2}\dots R_{c_{\rho(x)}}\underbrace{\cdots RTR}_{m(R,T)}$, which is linked to $R_{c_1}R_{c_2}\dots R_{c_{\rho(x)}}\underbrace{\cdots TRT}_{m(R,T)}$, which, again by the induction hypothesis, is linked to $T_1T_2\dots T_{\rho(z)}$. So we are done.

Suppose now that $R_1R_2\dots R_k$ is not reduced, and let i be such that $R_1R_2\dots R_i$ is reduced while $R_1R_2\dots R_iR_{i+1}$ is not. Consider $x := R_i\dots R_1(\hat{0})$. Then there exists a reduced expression for x ending with R_{i+1}. By what we have already proved, this expression is linked to $R_1R_2\dots R_i$ by a sequence of braid moves. Hence $R_1R_2\dots R_iR_{i+1}$ is linked to a reduced expression of length $i-1$ by a sequence of braid moves and a nil move. By iterating this proceeding, we have that $R_1R_2\dots R_k$ is linked to a reduced expression for z by a sequence of braid and nil moves. Then we get the assertion by what we have already proved. □

Remark. Any set $\mathcal{R}$ of special matchings of P satisfying properties 1 and 2 must be in bijection with the set of atoms of P. In fact, for all $s \rhd \hat{0}$, there exists a unique $R_s \in \mathcal{R}$ such that $R_s(s) = \hat{0}$. The existence is given by property 1 and the uniqueness follows by property 2 since if $R, R' \in \mathcal{R}$ coincide on an element, they must coincide everywhere.

We can now prove the main result of this work.

THEOREM 4.2. *A graded poset P is isomorphic to a Coxeter group W partially ordered by Bruhat order if and only if there exists a set $\mathcal{R}$ of special matchings of P such that*

1. *for all $z \in P \setminus \{\hat{0}\}$, there exists $R \in \mathcal{R}$ such that $R(z) \lhd z$;*
2. *for all $R, R' \in \mathcal{R}$, all orbits under the action of $\langle R, R'\rangle$ have the same cardinality (possibly ∞).*

Proof. Given a Coxeter system (W, S), let $\mathcal{R}$ be the set of all special matchings of W given by left multiplication by a generator. Then W endowed with the structure of a poset given by Bruhat order is a graded poset, and it is easy to show that properties 1 and 2 are satisfied.

Conversely, let P be a graded poset with rank function ρ, $\mathcal{R}$ be a set of special matchings of P satisfying properties 1 and 2 and S be the set of atoms of P. For all $s \in S$, let R_s be the unique special matching in $\mathcal{R}$ such that $R_s(s) = \hat{0}$ (see the above remark). Consider the Coxeter system (W, S) with Coxeter matrix $(m(s,t))_{t,s\in S}$, where $m(s,t) := m(R_s, R_t)$ is equal to the common cardinality of all orbits under the action of $\langle R_s, R_t\rangle$ divided by 2. By iteration of property 1, every element in P admits at least one (reduced) expression of special matchings in $\mathcal{R}$. We define a map $\Phi : P \to W$ in the following way. If $z \in P$ admits the expression $R_{s_1}R_{s_2}\dots R_{s_h}$, then $\Phi(z) = s_1s_2\dots s_h$. By Theorems 2.2, 2.3, 3.5, and 4.1, the map Φ is well defined, bijective, and order-preserving (as well as the inverse). □

Theorem 4.2 gives a characterization of the Coxeter group W partially ordered by Bruhat order among all posets. The following result gives a characterization of the Bruhat order among all partial orders on the set W.

THEOREM 4.3. *Let (W, S) be a Coxeter system. The Bruhat order provides the unique partial order structure on W for which the identity e is the bottom element and the multiplication on the left (equivalently, on the right) by s is a special matching for every $s \in S$.*

Proof. For the symmetry of the problem, we prove only the left version of the statement. Let $(W, \leq')$ be a partial order structure on W for which the multiplication on the left by s is a special matching for every $s \in S$, and denote by $\lhd'$ and $\rhd'$ its covering relations.

Suppose that $\{(x, y) : x \lhd y, x \not\lhd' y\} \neq \emptyset$. Let $x \lhd y$ be of minimal rank in that set, and let $s_1 \dots s_n$ be a reduced expression for y. It is well known that there exists a unique $r \in [n]$ such that $s_1 \dots s_{r-1} s_r s_{r-1} \dots s_1 y = x$. Let $u := s_{r+1} \dots s_n$. Then $s_r u \rhd u$, $s_{r-1} u \rhd u$, $s_{r-1} s_r u \rhd s_r u$. Hence, since the multiplication on the left by s_{r-1} is a special matching, we have $s_{r-1} s_r u \rhd s_{r-1} u$. Now, considering $s_{r-1} s_r u \rhd s_{r-1} u$ and the special matching given by multiplication on the left by s_{r-2}, we obtain $s_{r-2} s_{r-1} s_r u \rhd s_{r-2} s_{r-1} u$. By iteration, we have $s_1 y = s_2 \dots s_{r-1} s_r u \rhd s_2 \dots s_{r-1} u = s_1 x$, and hence $s_1 y \rhd' s_1 x$ by the minimality of the pair (x, y). Now, since the multiplication by s_1 is a special matching of $(W, \leq')$, it must be $y \rhd' x$. Hence $\{(x, y) : x \lhd y, x \not\lhd' y\} = \emptyset$.

Suppose now that $\{(x, y) : x \lhd' y, x \not\lhd y\} \neq \emptyset$. Let $x \lhd' y$ be of minimal rank in that set, and let $s \in D_L(y)$ (that is, $sy \lhd y$). By what we have already proved, $sy \lhd' y$. Since the multiplication by s is a special matching of $(W, \leq')$, $sx \lhd' x$ and $sx \lhd' sy$. By the minimality of (x, y), we have $sx \lhd x$ and $sx \lhd sy$. But, since the multiplication by s is a special matching of $(W, \leq)$, it must be $x \lhd y$. Hence $\{(x, y) : x \lhd' y, x \not\lhd y\} = \emptyset$, and we get the assertion. ∎

We end this section giving a corollary of Theorems 4.2 and 4.3 which shows how to reconstruct the weak Bruhat order from the (strong) Bruhat order and vice versa.

Let (W, S) be a Coxeter system, and let $\leq$ and $\leq_w$ denote, respectively, the (strong) Bruhat order and the weak Bruhat order. Given a poset $(P, \preceq)$ isomorphic to $(W, \leq)$, for every set $\mathcal{R}$ of special matchings of $(P, \preceq)$ satisfying the properties as in Theorem 4.2, we define another partial order $\preceq_{\mathcal{R}}$ on P as follows. For all $x \preceq y \in P$, y covers x in $(P, \preceq_{\mathcal{R}})$ if and only if there exists $R \in \mathcal{R}$ such that $R(x) = y$. On the other hand, given a poset $(P, \preceq)$ isomorphic to $(W, \leq_w)$, for every labeling L of the Hasse diagram of $(P, \preceq)$ satifying the properties as in Theorem 2.4, we define another partial order $\preceq_L$ on P as follows. For all $s \in S$, let $M_s : P \to P$ be the map sending x to y if and only if $L(\{x, y\}) = s$. Then set $\preceq_L$ to be the unique partial order on P such that M_s is a special matching for all $s \in S$. The existence of $\preceq_L$ is given by the poset $(W, \leq)$ identifying the elements of W with the elements of P via the poset isomorphism chosen according to the labeling L; the uniqueness part follows by Theorem 4.3. Now the following result follows by Theorem 4.2 (and by the definition of the poset isomorphism Φ appearing in its proof).

COROLLARY 4.4. *Let (W, S) be a Coxeter system, and let $\leq$ and $\leq_w$ denote, respectively, the (strong) Bruhat order and the weak Bruhat order.*

1. *Given a poset $(P, \preceq)$ isomorphic to $(W, \leq)$, the poset $(P, \preceq_{\mathcal{R}})$ is isomorphic to $(W, \leq_w)$ for all $\mathcal{R}$.*
2. *Given a poset $(P, \preceq)$ isomorphic to $(W, \leq_w)$, the poset $(P, \preceq_L)$ is isomorphic to $(W, \leq)$ for all L.*

5. Shellability for zircons. In this section we show how the results in section 3 imply that certain posets are CL-shellable (chain lexicographically shellable) in the sense of Björner and Wachs [6]. For several applications of CL-shellability, we refer the reader to [2], [3], [4].

Let P be a graded poset with $\hat{1}$. Let $\mathcal{M}$ be the set of maximal chains of P and $\mathcal{E} := \{(c, x, y) | c \in \mathcal{M}, x, y \in c, x \lhd y\}$. A *chain-edge labeling* of P is a map $\lambda : \mathcal{E} \to \mathbb{Z}$ such that if two maximal chains coincide along their last d edges, then their labels also coincide along these edges: if $c := \hat{0} = z_0 \lhd z_1 \lhd \cdots \lhd z_k = \hat{1}$, $c' := \hat{0} = z'_0 \lhd z'_1 \lhd \cdots \lhd z'_k =$

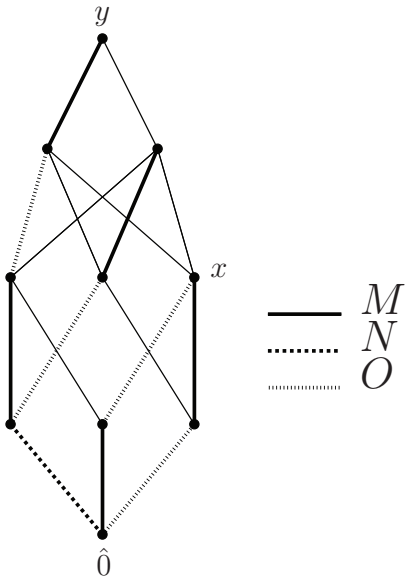


FIG. 3. *NMOM is not a B-regular expression.*

$\hat{1}$, and $z_h = z'_h$ for all $h \in [k-d, k]$, then $\lambda(c, z_h, z_{h+1}) = \lambda(c', z'_h, z'_{h+1})$ for all $h \in [k-d, k-1]$. For all maximal chains $c := \hat{0} = z_0 \lhd z_1 \lhd \cdots \lhd z_k = \hat{1}$, we let $\lambda(c) := \lambda(c, z_{k-1}, z_k)\lambda(c, z_{k-2}, z_{k-1}) \ldots \lambda(c, z_0, z_1)$.

A *rooted interval* with root r is a pair $([x, y], r)$, where $x \leq y \in P$ and r is a maximal chain in $[y, \hat{1}]$. If c is any maximal chain of $[x, y]$, then $c \cup r$ is a maximal chain of $[x, \hat{1}]$. Let λ be a chain-edge labeling of P. Once fixed the root r, by the definition of chain-edge labelings, λ restricts to a chain-edge labeling on $[x, y]$ since every maximal chain of P containing $c \cup r$ induces the same labeling on c. We denote the restriction map simply by λ_r when the interval $[x, y]$ is clear from the context.

A chain-edge labeling λ is said to be a CL-labeling if for every rooted interval $([x, y], r)$ in P

(i) there is a unique maximal chain c in $[x, y]$ such that $\lambda_r(c)$ is increasing and
(ii) for all other maximal chains c' in $[x, y]$, $\lambda_r(c) < \lambda_r(c')$ in the lexicographical order.

A graded poset P is said to be CL-shellable if every interval of P admits a CL-labeling.

Now let Z be a zircon, $x \leq y \in Z$, and $M_1 M_2 \ldots M_{\rho(y)}$ be a reduced expression of special matchings for y. We describe a chain-edge labeling λ on the interval $[x, y]$ depending on the reduced expression $M_1 M_2 \ldots M_{\rho(y)}$. Let $k := \rho(y) - \rho(x)$ and $c := x = z_0 \lhd z_1 \lhd \cdots \lhd z_k = y$ be a maximal chain of $[x, y]$. By Theorem 3.4, z_{k-1} admits a unique reduced expression which is obtainable from $M_1 M_2 \ldots M_{\rho(y)}$ deleting one letter M_i. Let $\lambda(c, z_{k-1}, y) := i$. Now consider the reduced expression $M_1 \ldots \widehat{M_i} \ldots M_\rho$ (M_i deleted) for z_{k-1}, and apply Theorem 3.4 to $z_{k-2} \lhd z_{k-1}$. If i' is such that the expression obtained from $M_1 M_2 \ldots M_{\rho(y)}$ by deleting both M_i and $M_{i'}$ is a reduced expression for z_{k-2}, then let $\lambda(c, z_{k-2}, z_{k-1}) := i'$. By iterating this procedure, we obtain a map λ which is evidently a chain-edge labeling.

In general, λ is not a CL-labeling. Consider the zircon in Figure 3 and the map λ induced on the interval $[x, y]$ by the reduced expression $NMOM$ for y. There is no chain c such that $\lambda_{\{y\}}(c)$ is increasing since the two chains are labeled 41 and 21.

We need the following definition appearing in [9, Definition 10.1] in the context of Coxeter groups.

DEFINITION 5.1. *We say that a reduced expression $M_1M_2\dots M_{\rho(y)}$ of special matchings for y is* B-regular *if*

$$M_i(x) \neq M_{i+1}M_{i+2}\dots M_{i+k}\dots M_{i+2}M_{i+1}(x)$$

for all $i \in [\rho(y)]$, $k \in [\rho(y)-i]$, and for all $x \in [\hat{0}, y]$ for which both sides are defined.

We say that a zircon Z is *B-regular* if every $y \in Z$ admits a reduced expression of special matchings whose reduced subwords are B-regular. The expression in Figure 3 is not B-regular since $M(\hat{0}) = OMO(\hat{0})$. Nevertheless, the zircon which is depicted is B-regular, a reduced expression for the top element whose reduced subwords are B-regular being the one in Figure 1.

Thanks to the results in the previous sections, we can prove the following result.

COROLLARY 5.2. *Every B-regular zircon Z is a CL-shellable poset.*

Proof. Let ρ be the rank function of Z, and consider an interval $[x, y]$ in Z. Let $M_1M_2\dots M_{\rho(y)}$ be a reduced expression for y whose reduced subwords are B-regular, and let λ be the chain-edge labeling induced by it. By our assumption on $M_1M_2\dots M_{\rho(y)}$, no generality is lost if we verify (i) and (ii) only for the trivial rooted interval $([x, y], \{y\})$ instead of considering all rooted intervals in $[x, y]$.

To show that two distinct maximal chains in $[x, y]$ cannot both have increasing labels, we proceed by contradiction. Let $[x, y]$ be of minimal rank $k := \rho(y) - \rho(x)$ among the intervals that admit two distinct maximal chains with increasing labels. Let c and c' be two distinct maximal chains of $[x, y]$ both with increasing labels $\lambda(c) = (i_1, i_2, \dots, i_k)$ and $\lambda(c') = (i'_1, i'_2, \dots, i'_k)$. Assume that $i_k \leq i'_k$. The element x admits the reduced expression $M_1M_2\dots\widehat{M_{i_1}}\dots\widehat{M_{i_k}}\dots M_{\rho(y)}$. If $i_k < i'_k$, then $M_{\rho(y)}M_{\rho(y)-1}\dots M_{i'_k+1}M_{i'_k}M_{i'_k+1}\dots M_{\rho(y)-1}M_{\rho(y)}(x) < x$, but this is a contradiction since $M_{\rho(y)}M_{\rho(y)-1}\dots M_{i'_k+1}M_{i'_k}M_{i'_k+1}\dots M_{\rho(y)-1}M_{\rho(y)}(x) \in c'$. If $i_k = i'_k$, the contradiction stems from the minimality of the interval $[x, y]$ since the element $M_{\rho(y)}M_{\rho(y)-1}\dots M_{i_k+1}M_{i_k}M_{i_k+1}\dots M_{\rho(y)-1}M_{\rho(y)}(x)$ belongs to $\in c \cap c'$.

It remains to prove that there exists a maximal chain c with increasing labels and that $\lambda(c) < \lambda(c')$ in the lexicographical order for any other maximal chain c'. By Theorem 3.4, we can construct the chain $c := x = z_0 \lhd z_1 \lhd \cdots z_k = y$ in the following way. Every coatom of y in $[x, y]$ admits a reduced expression which is obtained from $M_1M_2\dots M_{\rho(y)}$ by deleting one letter. Among these elements, let z_{k-1} be the coatom whose reduced expression is obtained deleting the letter M_i with i minimal. Now repeat the procedure considering the reduced expression $M_1M_2\dots\widehat{M_i}\dots M_{\rho(y)}$ for z_{k-1} together with the coatoms of z_{k-1} in $[x, z_{k-1}]$ and so on. By construction, $\lambda(c) < \lambda(c')$ in the lexicographical order for any other maximal chain c'. We need to show that $\lambda(c) = (i_1, i_2, \dots, i_k)$ is increasing. Suppose by contradiction that it is not, and let r be such that $i_r > i_{r+1}$. This means that our procedure fails to give a chain with increasing labels in the rank 2 interval $[x', y']$, where $x' := M_1M_2\dots\widehat{M_{i_1}}\dots\widehat{M_{i_{r+1}}}\dots M_\rho(y)$ (with $r+1$ letters deleted) and $y' := M_1M_2\dots\widehat{M_{i_1}}\dots\widehat{M_{i_{r-1}}}\dots M_\rho(y)$ (with $r-1$ letters deleted). So it remains to prove that the assertion holds for the rank 2 intervals.

From now on assume that $\rho(y) - \rho(x) = 2$, and recall from Theorem 2.8 that the interval $[x, y]$ is a square. Let (i, j) be lexicographically minimal among the ordered pairs (h, k) such that $M_1M_2\cdots\widehat{M_h}\dots\widehat{M_k}\dots M_{\rho(y)}$ is a reduced expression for x. Let $u := M_{\rho(y)}M_{\rho(y)-1}\dots\widehat{M_i}\dots M_1(\hat{0})$. We claim that $u \in [x, y]$. By Theorem 3.5, $u < y$, and $u > x$ follows if we show that the expression $M_1M_2\dots\widehat{M_i}\dots M_{\rho(y)}$ is reduced.

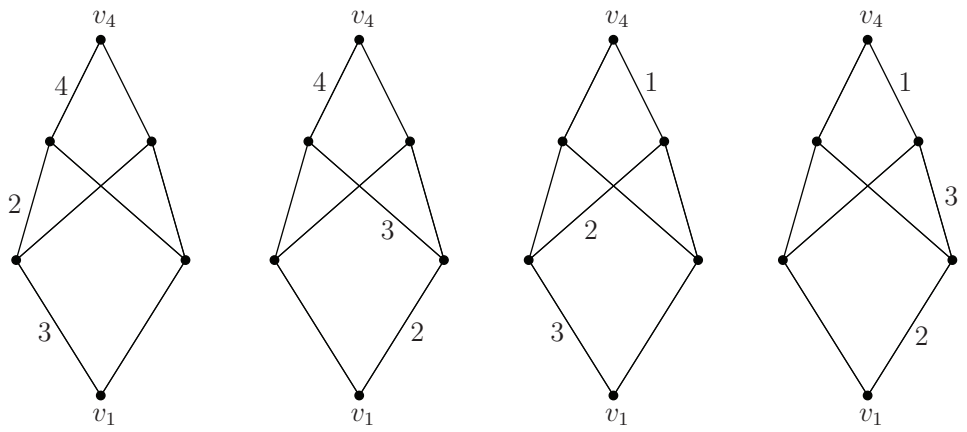


FIG. 4. *The CL-labeling on* $[v_1, v_4]$ *induced by* $MONM$.

Clearly $M_1M_2 \dots \widehat{M_i} \dots M_{j-1}$ is reduced. Let $v := M_{j-1}M_{j-2} \dots \widehat{M_i} \dots M_1(\hat{0})$. If $M_j(v) \lhd v$, then by Theorem 3.4 there exists $r \in [j-1]$ such that

$$M_j(v) = M_jM_{j-1} \dots \widehat{M_i} \dots M_1(\hat{0}) = M_{j-1} \dots \widehat{M_r} \dots \widehat{M_i} \dots M_1(\hat{0})$$

(note that r can be $< i$). Hence $v = M_jM_{j-1} \dots \widehat{M_r} \dots \widehat{M_i} \dots M_1(\hat{0})$, and the expression $M_1M_2 \dots \widehat{M_i} \dots \widehat{M_r} \dots M_{\rho(y)}$ for x is reduced. This contradicts the minimality of (i, j).

Then $M_j(v) \rhd v$, and $M_1M_2 \cdots \widehat{M_i} \dots M_j$ is reduced. We have $M_j(v) \neq M_{j+1}(v)$ since the expression $M_1M_2 \dots M_{\rho(y)}$ is B-regular. By the definition of a special matching, $M_{j+1}M_j(v) \rhd M_j(v)$ and $M_{j+1}M_j(v) \rhd M_{j+1}(v)$ since $M_{j+1}(v) \rhd v$, and so also the expression $M_1M_2 \dots \widehat{M_i} \dots M_{j+1}$ is reduced. Similarly, $M_{j+1}M_j(v) \neq M_{j+2}M_{j+1}(v)$ since the expression $M_1M_2 \dots M_{\rho(y)}$ is B-regular. Hence $M_{j+2}M_{j+1}M_j(v) \rhd M_{j+1}M_j(v)$ and $M_{j+2}M_{j+1}M_j(v) \rhd M_{j+2}M_{j+1}(v)$. Hence $M_1M_2 \dots \widehat{M_i} \dots M_{j+1}$ is reduced. By iteration, $M_1M_2 \dots \widehat{M_i} \dots M_{\rho(y)}$ is reduced, and we get the claim.

From the claim, it follows that $c = x \lhd u \lhd y$ and that $\lambda(c) = (i, j)$. Since $i < j$, $\lambda(c)$ is increasing. This completes the proof. ∎

Every Coxeter group is a B-regular zircon since the expressions of special matchings coming from reduced expressions in the sense of Coxeter group theory are B-regular (see the first remark in section 3). Corollary 5.2 is inspired by, and generalizes, the analogous result for Coxeter groups which has been proven by Björner and Wachs [5]. Here special matchings play the role that Coxeter generators play in Björner–Wachs' proof.

Example. Consider the B-regular zircon in Figure 1, its interval $[v_1, v_4]$, and the reduced expression $MONM$ of special matchings for v_4. All reduced subwords of $MONM$ are B-regular. The induced CL-labeling λ on $[v_1, v_4]$ is as in Figure 4.

REFERENCES

[1] A. BJÖRNER AND F. BRENTI, *Combinatorics of Coxeter Groups*, Grad. Texts in Math. 231, Springer, New York, 2005.

[2] A. BJÖRNER, *Shellable and Cohen-Macaulay partially ordered sets*, Trans. Amer. Math. Soc., 1 (1980), pp. 159–183.

[3] A. Björner, *Posets, regular CW complexes and Bruhat order*, European J. Combin., 5 (1984), pp. 7–16.
[4] A. Björner, A. Garsia, and R.P. Stanley, *An introduction to Cohen-Macaulay partially ordered sets*, in Ordered Sets, Reidel, Dordrecht/Boston, 1982, pp. 583–615.
[5] A. Björner and M. Wachs, *Bruhat order of Coxeter groups and shellability*, Adv. Math., 43 (1982), pp. 87–100.
[6] A. Björner and M. Wachs, *On lexicographically shellable posets*, Trans. Amer. Math. Soc., 277 (1983), pp. 323–341.
[7] N. Bourbaki, *Groupes et Algèbres de Lie*, Chs. 4–6, Hermann, Paris, 1968.
[8] F. Brenti, *The intersection cohomology of Schubert varieties is a combinatorial invariant*, European J. Combin., 25 (2004), pp. 1151–1167.
[9] F. Brenti, F. Caselli, and M. Marietti, *Special matchings and Kazhdan-Lusztig polynomials*, Adv. Math., 202 (2006), pp. 555–601.
[10] V.V. Doedhar, *Some characterizations of Bruhat ordering on a Coxeter group and determination of the relative Möbius function*, Invent. Math., 39 (1977), pp. 187–198.
[11] V.V. Doedhar, *Some characterizations of Coxeter groups*, Enseign. Math., 32 (1986), pp. 111–120.
[12] K. Eriksson, *Polygon posets and the weak order of Coxeter groups*, J. Algebraic Combin., 4 (1995), pp. 233–252.
[13] D.M. Goldschmidt, *Abstract reflections and Coxetere groups*, Trans. Amer. Math. Soc., 67 (1977), pp. 209–214.
[14] H. Hiller, *Geometry of Coxeter Groups*, Res. Notes Math. 54, Pitman Advanced Publishing Program, Boston, 1982.
[15] J.E. Humphreys, *Reflection Groups and Coxeter Groups*, Cambridge Stud. Adv. Math. 29, Cambridge University Press, Cambridge, 1990.
[16] M. Marietti, *Kazhdan-Lusztig Theory: Boolean Elements, Special Matchings and Combinatorial Invariance*, Ph.D. thesis, Università degli Studi di Roma La Sapienza, Roma, Italy, 2003.
[17] M. Marietti, *Algebraic and combinatorial properties of zircons*, J. Algebraic Combin., 26 (2007), pp. 363–382.
[18] H. Matsumoto, *Générateurs et relations des groupes de Weyl généralisés*, C. R. Acad. Sci. Paris, 258 (1964), pp. 3419–3422.
[19] J. Tits, *Le problème des mots dans les groupes de Coxeter*, in Symposia Mathematica, Vol. 1, Academic Press, London, 1969, pp. 175–185.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 333–343

ON CUSICK'S METHOD AND VALUE SETS OF CERTAIN POLYNOMIALS OVER FINITE FIELDS*

PETRI ROSENDAHL†

Abstract. In this paper, we consider Cusick's method to find the number of values of the polynomials $f_a(x) = x^a\,(x+1)^{2^k-1}$, when $x \in GF(2^{2k})$. We will prove that under certain conditions $f_a(x)$ and $f_{2-a}(x)$ have the same number of values. We will also prove a conjecture due to Cusick.

Key words. finite fields, value sets, cross-correlation functions, Niho type decimations

AMS subject classifications. 11T06, 11T23, 94B15, 94A55

DOI. 10.1137/050626570

1. Introduction. We will denote the finite field with p^n elements by $GF(p^n)$, and its multiplicative group is denoted by $GF(p^n)^\times$. In this paper $p = 2$ except in section 2.3. We will assume that the reader has a basic knowledge of finite fields. Especially, the reader should be familiar with the trace function tr: $GF(p^n) \to GF(p)$,

$$\mathrm{tr}(x) = x + x^p + x^{p^2} + \cdots + x^{p^{n-1}},$$

and the canonical additive character $\chi : GF(p^n) \to \mathbb{C}^\times$, which is defined by

$$\chi(x) = \zeta^{tr(x)},$$

where $\zeta = e^{2\pi i/p}$.

We will also assume that the reader is familiar with the cross-correlation function between two m-sequences; this essentially amounts to the theory of the character sum

$$C_d(y) = \sum_{x \in GF(p^n)} \chi(yx + x^d).$$

In the following, n will always be even, say $n = 2k$. The conjugate of an element $y \in GF(p^n)$ over $GF(p^k)$ will be denoted by $\overline{y}$, i.e.,

$$\overline{y} = y^{p^k}.$$

The group of $\big(p^k + 1\big)$st roots of unity in $GF(p^n)$ will be denoted by S, that is,

$$S = \{x \in GF(p^n) \mid x\overline{x} = 1\}.$$

The conjugation operation has many properties similar to the usual complex conjugation. For example, we have $\overline{x+y} = \overline{x} + \overline{y}$ and $\overline{xy} = \overline{x}\,\overline{y}$ for all $x, y \in GF(p^n)$. We also have $\overline{x} = x$ for $x \in GF(p^k)$ and $\overline{x} = x^{-1}$ for $x \in S$.

In this paper we will study the value sets

$$V(f_a) = \{f_a(x) \mid x \in GF(2^n)\}$$

*Received by the editors March 11, 2005; accepted for publication (in revised form) September 15, 2008; published electronically January 7, 2009.
http://www.siam.org/journals/sidma/23-1/62657.html
†Department of Mathematics, University of Turku, 20014 Turku, Finland (perosen@utu.fi).

of the polynomials of the form

$$f_a(x) = x^a \left(x+1\right)^{2^k-1}.$$

The study of the value sets of the polynomials $f_a(x)$ was originated by Cusick in [2], and this paper is a continuation of Cusick's studies. We will formulate Cusick's approach in its general setting. Furthermore, we will prove that exponents a and $a-2$ will give value sets of the same cardinality. Lastly, we prove a conjecture due to Cusick.

The fact that the cardinality of a value set can be counted is interesting in itself. This is because, in general, apart from the case of permutation polynomials, there are very few results on the value sets of polynomials over finite fields. Moreover, the polynomials of the above form are of special interest because of their connection to cross-correlation functions; see [2]. Also, in [3] it was shown that for $a = 1$ the number of values of $f_a(x)$ attains a bound due to Wan [12].

An integer (or a decimation or an exponent) d satisfying

$$d \equiv 1 \pmod{p^k - 1},$$

is said to be of Niho type. Also, the corresponding power functions $x \mapsto x^d$ in $GF(p^n)$ and cross-correlation functions of the m-sequences of period $p^n - 1$ are said to be of Niho type.

Power functions and cross-correlation functions corresponding to Niho-type exponents have attained a lot of interest in the past few years. First, in [8] a family of four-valued cross-correlation functions of m-sequences was found. Second, in [1] it was proven that Niho-type cross-correlation functions have at least four different values. Third, in [7] the equation $(x+1)^d = x^d + 1$ was studied for Niho type d; this is of use in finding the distribution of values of cross-correlation functions and weight distributions of certain cyclic codes. Lastly, we mention that properties of Niho-type power functions were exploited in the construction of bent functions in [5].

2. Cusick's method.

2.1. Value sets and cross-correlation functions. In this section we review the main method used in [2]. Cusick uses this approach ad hoc; here we have tried to formulate this method in a more general setting. We restrict ourselves to the binary case. We will give an example of the nonbinary case in section 2.3.

First of all, let

$$f_a(x) = x^a \left(x+1\right)^{2^k-1},$$

where a is an arbitrary integer. Clearly, if $a \equiv b \pmod{2^n - 1}$, then $f_a(x)$ and $f_b(x)$ represent the same polynomial functions in $GF(2^n)$. As usual, we consider x^{-1} and x^{2^n-2} as the same functions and therefore always $f_a(0) = 0$.

We wish to find $|V(f_a)|$, i.e., the number of values of f_a, when $x \in GF(2^n)$, where $n = 2k$.

To begin with, we give the following two lemmas. Recall that

$$S = \{x \in GF(2^{2k}) \mid x\overline{x} = 1\},$$

where $\overline{x} = x^{2^k}$.

LEMMA 2.1. *Assume* $\gcd(a, 2^k - 1) = 1$. *Then* $f_a(\gamma) \in S$ *if and only if* $\gamma \in S \setminus \{1\}$.

Proof. Clearly, $(\gamma+1)^{2^k-1} \in S$ for all $\gamma \neq 1$. Therefore, if $\gamma \in S \setminus \{1\}$, then obviously $f_a(\gamma) \in S$. So assume $\gamma^a (\gamma+1)^{2^k-1} \in S$. Then also $\gamma^a \in S$ since S is a subgroup of $GF(2^{2k})^\times$ and $(\gamma+1)^{2^k-1} \in S$. Thus $\gamma^{a(2^k+1)} = 1$. We have $\gamma^{2^k+1} \in GF(2^k)^\times$, and because by assumption the equation $x^a = 1$ has a unique solution in $GF(2^k)^\times$, we must have $\gamma^{2^k+1} = 1$. ∎

We will also need the following fact, which is also noted in [2].

LEMMA 2.2. *If $\beta \in S \setminus \{1\}$, then*

$$f_a(\beta) = \beta^{a-1}.$$

Proof. The claim follows easily from the fact that for $\beta \in S \setminus \{1\}$ we have $\beta\overline{\beta} = 1$, and therefore

$$\beta (\beta+1)^{2^k-1} = \frac{\beta\left(\overline{\beta}+1\right)}{(\beta+1)} = \frac{(1+\beta)}{(\beta+1)} = 1.$$

Note that the lemma is valid in fields of odd characteristic also, when $S \setminus \{1\}$ is changed to $S \setminus \{-1\}$. ∎

From now on we will assume that $\gcd(a, 2^n - 1) = 1$. Since $n = 2k$ this implies $\gcd(a, 2^k - 1) = 1$.

Cusick's method is as follows. Assume that we know the number of $y \in GF(2^{2k})$ such that the equation

$$x^a + yx^{a-1} + \overline{y}x + 1 = 0 \tag{2.1}$$

has a solution $x \in S$.

Let $\beta \in S$ be a solution to (2.1), and define γ by $y = \beta\gamma$. For a moment we assume that $y \notin S$, and then we have $\gamma \neq 1$. Then (note that $\overline{\beta} = 1/\beta$)

$$\beta^a + \gamma\beta^a + \overline{\gamma} + 1 = 0,$$

so we have that

$$\beta^a = (\gamma+1)^{2^k-1}.$$

Therefore

$$y^a = \gamma^a (\gamma+1)^{2^k-1}, \tag{2.2}$$

that is, $y^a \in V(f_a)$.

Conversely, let $\gamma \neq 1$ be given, and define y by

$$y^a = \gamma^a (\gamma+1)^{2^k-1}.$$

Furthermore, let $\beta \in S$ be given by

$$\beta^a = (\gamma+1)^{2^k-1}.$$

The assumption $\gcd(a, 2^{2k}-1) = 1$ now implies $y = \gamma\beta$. Therefore

$$\begin{aligned}\beta^a + y\beta^{a-1} + \overline{y}\beta + 1 &= (\gamma+1)\,\beta^a + (\overline{\gamma}+1)\\ &= (1+\gamma)^{2^k} + (\overline{\gamma}+1)\\ &= 0,\end{aligned}$$

i.e., β is a solution to (2.1).

Because of Lemma 2.1, we have now established a one-to-one correspondence between the sets

$$\{y \in GF(2^n) \mid y \notin S,\ (2.1) \text{ has a solution in } S\}$$

and

$$\{x \in V(f_a) \mid x \notin S\}.$$

We still have to consider for which $x \in S$ we have $x \in V(f_a)$.

Lemma 2.2 implies that $f_a(S)$ consists of the 0 and the elements β^{a-1}, where $\beta \in S \setminus \{1\}$. Therefore,

$$|f_a(S)| = 2^k + 1,$$

if $\gcd\left(a-1, 2^k+1\right) = 1$. If $\gcd\left(a-1, 2^k+1\right) > 1$, then we have

$$|f_a(S)| = \frac{2^k+1}{\gcd(a-1, 2^k+1)} + 1.$$

All in all, we have the following theorem. Note that if $y \in S$, then (2.1) always has a solution in S, since in this case the equation factors as

$$\left(x^{a-1} + \overline{y}\right)(x+y).$$

Note also that $f_a(0) = 0$ has to be taken into account.

THEOREM 2.3. *Denote by N the number of $y \in GF(2^n)$ such that (2.1) has a solution in S. Moreover, assume that $\gcd(a, 2^n-1) = 1$. Then*

$$|V(f_a)| = N - 2^k - 2 + |f_a(S)|.$$

Hence, if $\gcd\left(a-1, 2^k+1\right) = 1$, then we have

$$|V(f_a)| = N - 1,$$

and if $\gcd\left(a-1, 2^k+1\right) > 1$, then we have

$$|V(f_a)| = N - 2^k - 1 + \frac{2^k+1}{\gcd(a-1, 2^k+1)}.$$

We still have to study (2.1). The following classic theorem will be crucial.

THEOREM 2.4 (Niho's theorem). *Assume that d is of Niho type, and define s by $d = (p^k-1)\cdot s + 1$. Then, as y ranges over $GF(p^n)$, the values of the character sum*

$$\sum_{x \in GF(p^n)} \chi\left(yx + x^d\right) \tag{2.3}$$

are exactly the same as the values

$$(N(y) - 1) \cdot p^k,$$

where $N(y)$ *is the number of* $x \in S$ *such that*

$$x^{2s-1} + yx^s + \overline{y}x^{s-1} + 1 = 0. \tag{2.4}$$

This theorem was proved by Niho in his thesis [9] for $p = 2$. Proofs for this can be found also, e.g., in [4], [2], and [10]. The generalization for all primes p is from the author; see [10, 11].

It is noted in [8] (see also [4]) that if $\gcd(t, p^k+1) = 1$, then x^t may be substituted in place of x in (2.4). This may give a more tractable equation.

We will now show how (2.4) can be transformed into the form (2.1). Assume that $\gcd(s-1, 2^k+1) = 1$ so $t = (s-1)^{-1} \pmod{2^k+1}$ exists. Then we can substitute x^t in place of x in (2.4) to get

$$x^{t+2} + yx^{t+1} + \overline{y}x + 1 = 0. \tag{2.5}$$

Note that this equation is of the form (2.1) (for $a = t + 2$).

On the other hand, assuming that $\gcd(a-2, 2^k+1) = 1$, from (2.4) we get (2.1) by substituting x^{s-1} for x, where s is defined by $(a-2)^{-1} = s - 1 \pmod{2^k+1}$.

The previous remarks give the connection between the character sum (2.3) and the equation (2.1). That is, when $d = (2^k - 1) \cdot s + 1$ and $\gcd(s-1, 2^k+1) = 1$, then the nonnegative values of the sum (2.3) are connected to the values of f_a, where $a = (s-1)^{-1} + 2$ via Theorem 2.3.

For later use, we note the following:

$$\gcd(d, 2^k+1) = \gcd(t+2, 2^k+1) = \gcd(2s-1, 2^k+1),$$

where d, s, and t are as above. We have assumed that $\gcd(d, 2^n - 1) = 1$, and this is always the case in the context of cross-correlation functions of m-sequences. In section 3, we treat a case where $\gcd(d, 2^n - 1) > 1$.

2.2. Equivalence of certain exponents. The results in [2] on the cardinalities of value sets seem to come in pairs. More precisely, the cardinalities of value sets of f_a and f_{2-a} seem to be the same assuming certain conditions on greatest common divisors. In the following we will see that this indeed is the case.

Assume that $\gcd(d, 2^n - 1) = 1$ and consider the character sums

$$\sum_{x \in GF(2^n)} \chi\left(yx + x^d\right) \tag{2.6}$$

and

$$\sum_{x \in GF(2^n)} \chi\left(x + yx^d\right). \tag{2.7}$$

It is fairly easy to see that the values and distribution of values are exactly the same for both sums (however, it is not true that they always have the same value for the same y). More precisely,

$$\sum_{x \in GF(2^n)} \chi\left(x + yx^d\right) = \sum_{z \in GF(2^n)} \chi\left(bz + z^d\right),$$

where $b = y^{-1/d}$. One may think of the previous sums as values of cross-correlation functions of m-sequences $\mathrm{tr}(\alpha^i)$ and $\mathrm{tr}(\alpha^{di})$, $i = 0, 1, 2, \ldots$, and then y corresponds to a cyclic shift of one of the sequences. Obviously, when y ranges over $GF(2^n)^\times$ it does not matter which one of the sequences is shifted.[1]

Assume that d is of Niho type, i.e., $d = (2^k - 1) \cdot s + 1$, and assume also that $\gcd(d, 2^n - 1) = 1$. The proof of Niho's theorem can be imitated for sum $\sum \chi(x + yx^d)$ assuming, of course, that d is of Niho type. So while the sum (2.6) leads to

$$x^{2s-1} + yx^s + \overline{y}x^{s-1} + 1 = 0, \tag{2.8}$$

the sum (2.7) leads to

$$yx^{2s-1} + x^s + x^{s-1} + \overline{y} = 0, \tag{2.9}$$

and in both equations we are interested in the number of roots in S.

As the values and their distribution are the same for both sums, we must have that (2.8) and (2.9) have identical patterns of solutions. More precisely, for each i the numbers

$$N_i = |\{y \in GF(2^n) \mid (2.8) \text{ has exactly } i \text{ solutions in } S\}|$$

and

$$N_i' = |\{y \in GF(2^n) \mid (2.9) \text{ has exactly } i \text{ solutions in } S\}|,$$

are the same.

Assume that $\gcd(s - 1, 2^k + 1) = 1$ so $t = (s - 1)^{-1} \pmod{2^k + 1}$ exists. Then we can substitute x^t in place of x in (2.8) and (2.9) to get

$$x^{t+2} + yx^{t+1} + \overline{y}x + 1 = 0, \tag{2.10}$$

$$yx^{t+2} + x^{t+1} + x + \overline{y} = 0, \tag{2.11}$$

Recall that N is the number of $y \in GF(2^n)$ such that (2.1) has a solution in S. We have $N = 2^n - N_0$. As we have seen, the number N (and hence the number N_0) determines the number of values of f_a, where $a = t + 2$.

Respectively, assume that $y \notin S$ and that $\beta \in S$ is a solution to (2.11), and define γ by $y = \beta^{-1}\gamma$. Then from (2.11) we get

$$\beta^{a-2} = (\gamma + 1)^{2^k - 1},$$

and therefore

$$y^{2-a} = \gamma^{2-a} (\gamma + 1)^{2^k - 1},$$

which says that $y^{2-a} \in V(f_{2-a})$.

Conversely, if $\gamma \neq 1$ is given, then we may define y using the previous equation, and proceed similarly as with the case of (2.8). Of course, to make the correspondence between the images y^{2-a} and (2.11) with a solution one to one, we have to assume that $\gcd(a - 2, 2^n - 1) = 1$.

[1] The case $y = 0$ is trivial.

Also the number of values of f_a and f_{2-a} in S are the same, because f_a restricted to $S \setminus \{1\}$ is x^{a-1} and f_{2-a} restricted to $S \setminus \{1\}$ is x^{1-a}; see Lemmas 2.1 and 2.2.

We formulate these observations as a theorem. Note that the given conditions guarantee that the substitution of x by x^t can be made, since we necessarily have $\gcd\left(a-2, 2^k+1\right) = 1$ also.

THEOREM 2.5. *Assume that* $\gcd(a, 2^n-1) = \gcd(2-a, 2^n-1) = 1$. *Then*

$$|V(f_a)| = |V(f_{2-a})|.$$

Remark 2.6. There are also some more or less trivially equivalent pairs of exponents, which are obtained by a change of the variable. For example, substituting x^{-1} for x in $x^a (x+1)^{2^k-1}$ gives the polynomial $x^{-a-2^k+1}(x+1)^{2^k-1}$. This must have a value set of the same size, since $x \mapsto x^{-1}$ is one to one (using convention $0^{-1} = 0$).

2.3. A nonbinary case. We have considered here the case $p = 2$ only, but some results make sense also in the case of odd characteristic. However, much less is known about Niho-type cross-correlation functions of nonbinary m-sequences. On the other hand, the methods presented in Cusick's paper [2] and here may give some information on the cross-correlation functions themselves. We will assume $p > 2$ throughout this section.

In [3] the number of values of $f(x) = x(x+1)^{q-1}$ in any finite extension of $GF(q)$ was found. For the extension $GF(q^2)$ this was done in a different way in [2]; the proof also applies word for word in the case $p > 2$.

We treat here the case

$$f(x) = x^3 (x+1)^{p^k-1}.$$

This corresponds to a cross-correlation function (character sum $\sum \chi\left(yx + x^d\right)$) of m-sequences which differ by the decimation $d = 2 \cdot p^k - 1$. Unfortunately, this d is the only Niho-type decimation for which the cross-correlation problem is completely solved. This is done in Theorem 4.13 of [6], from which the next lemma follows.

LEMMA 2.7. *Let* $n = 2k$, *and assume that* $p^k \not\equiv 2 \pmod 3$. *Then the number of* $y \in GF(p^n)$ *such that the equation*

$$x^3 + yx^2 + \overline{y}x + 1 = 0 \tag{2.12}$$

has a solution $x \in S$ *is*

$$N = \frac{1}{3}\left(2 \cdot p^{2k} + p^k\right).$$

THEOREM 2.8. *Assume that* $p^k \not\equiv 2 \pmod 3$. *The number of values of* $f(x) = x^3 (x+1)^{p^k-1}$ *in* $GF(p^{2k})$ *is*

$$\frac{1}{6}\left(4p^{2k} - p^k - 3\right).$$

Proof. We give a sketch of the proof only, since it is essentially same as in the binary case.

First, $f(x) \in S$ if and only if $x \in S$. Second, a one-to-one correspondence between the elements y such that (2.12) has a solution in S and the values $f(x)$ such that $x \notin S$ can be found by substituting $y = \gamma\beta$, with $\beta \in S$ a solution to (2.12). Finally, one has to consider the values of $f(x)$ in S. This is easily done since $f(x) = x^2$ when restricted to $S \setminus \{-1\}$. Note that for $p > 2$ we have that $f(x) = f(-x)$ implies that $f(x)$ takes on half of the values in S. ∎

3. A proof of Cusick's conjecture. Theorem 3.3 was conjectured in [2]. The crucial difference here is that when k is odd, we then have that $\gcd(d, 2^n - 1) = 3$, for d in the character sum (2.3) which leads to desired equation. Therefore d is not a decimation corresponding to a cross-correlation function of m-sequences, and we cannot directly apply results on these. However, the proof of Niho's theorem does not involve the condition $\gcd(d, 2^n - 1) = 1$ and we can still exploit the equation.

We will prove Theorem 3.1, which is a stronger result than Theorem 3.3. Theorem 3.1 is conjectured in [2] as Conjecture 1.

Let k be odd,

$$d = 2^{k+1} - 1,$$

and consider the character sums

$$\Delta_d(y) = \sum_{x \in GF(2^n)} \chi\left(x + yx^d\right) \tag{3.1}$$

and

$$\Delta'_d(y) = \sum_{x \in GF(2^n)} \chi\left(yx + x^d\right). \tag{3.2}$$

THEOREM 3.1. *The values of the character sum* (3.1) *are as follows:*

$$\begin{array}{llll} -2^k & occurs & \frac{1}{3}\left(2^{2k} - 2^k - 2\right) & times, \\ 0 & occurs & 2^{2k-1} - 2^{k-1} + 1 & times, \\ 2^k & occurs & 2^k & times, \\ 2^{k+1} & occurs & \frac{1}{3}\left(2^{2k-1} - 2^{k-1} - 1\right) & times, \end{array}$$

and the values of the sum (3.2) *are as follows:*

$$\begin{array}{llll} -2^k & occurs & \frac{1}{3}\left(2^{2k} - 2^k - 2\right) & times, \\ 0 & occurs & 2^{2k-1} - 2^{k-1} + 2 & times, \\ 2^k & occurs & 2^k - 2 & times, \\ 2^{k+1} & occurs & \frac{1}{3}\left(2^{2k-1} - 2^{k-1} + 2\right) & times. \end{array}$$

Proof. We treat the case of the sum (3.2) first. We have $d = 2 \cdot \left(2^k - 1\right) + 1$. By Niho's theorem, the values of the sum (3.2) are

$$(N(y) - 1) \cdot 2^k,$$

where $N(y)$ is the number of $x \in S$ such that

$$x^3 + yx^2 + \overline{y}x + 1 = 0. \tag{3.3}$$

Let N_i be the number of $y \in GF(2^n)$ such that (3.3) has exactly i distinct solutions in S. We will first find the number N_2.

Clearly, if two solutions of (3.3) are in S, then the third solution is also. Hence (3.3) can have exactly two solutions in S if and only if the corresponding polynomial has a double root and another root. The usual derivative argument shows that $x^2 = \overline{y}$,

and substituting this into (3.3) we get $y\overline{y} = 1$, i.e., necessarily $y \in S$. Then (3.3) splits as

$$\left(x^2 + \overline{y}\right)(x + y) = 0.$$

This has two distinct roots (namely $\sqrt{\overline{y}}$ and y) in S when $y \notin GF(4)$. Thus we have shown $N_2 = 2^k - 2$.

Second, we will need the following facts:

(i) $\sum N_i = 2^n$,
(ii) $\sum_{y \in GF(2^n)} \Delta'_d(y) = 2^n$,
(iii) $\sum_{y \in GF(2^n)} \Delta'_d(y)^2 = 2^{2n}$.

The fact (i) follows from the number of equations of the form (3.3). The equations (ii) and (iii) are well-known power sum identities, and are valid despite the fact $\gcd(d, 2^n - 1) > 1$. We now have the following system of linear equations:

$$\begin{aligned} N_2 &= 2^k - 2, \\ N_0 + N_1 + N_2 + N_3 &= 2^{2k}, \\ -2^k N_0 + 2^k N_2 + 2^{k+1} N_3 &= 2^{2k}, \\ 2^{2k} N_0 + 2^{2k} N_2 + 2^{2k+2} N_3 &= 2^{4k}. \end{aligned}$$

The claimed distribution for the sum (3.2) can now be calculated easily by solving the system.

We now turn to consider the sum (3.2). This time the values of the sum are

$$(N(y) - 1) \cdot 2^k,$$

where $N(y)$ is the number of $x \in S$ such that

$$yx^3 + x^2 + x + \overline{y} = 0. \tag{3.4}$$

As above, the derivative argument shows that if (3.4) has exactly two solutions in S, then necessarily $y \in S$. In this case (3.4) splits as

$$\left(yx^2 + 1\right)(x + \overline{y}) = 0.$$

Solutions to this are $x = \overline{y}$ and $x = \sqrt{y^{-1}}$. These are distinct when $y \neq 1$.

Let N'_i be the number of $y \in GF(2^n)$ such that (3.4) has exactly i solutions in S. We have shown that $N'_2 = 2^k$. In addition, we have (i) and (ii) above. However, (iii) now has the form

$$\sum_{y \in GF(2^n)} \Delta'_d(y)^2 = 2^{2n} - 2^{n+1}.$$

This is a special case of Lemma 3.2 below.

The rest of the proof is identical with the case of the sum Δ'_d. □

To fill the remaining gap, we still have to prove the following lemma.

LEMMA 3.2. *For all n, we have*

$$\sum_{y \in GF(2^n)} \left(\sum_{x \in GF(2^n)} \chi\left(x + yx^d\right) \right)^2 = 2^{2n} - 2^n \cdot (r - 1),$$

where $r = \gcd(d, 2^n - 1)$.

Proof. Unless otherwise stated, the range of the variables x, y, z, and α below is $GF(2^n)$; we only denote the restrictions when needed. We have

$$\begin{aligned}
\sum_y \left(\sum_x \chi\left(x + yx^d\right) \right)^2 &= \sum_y \left(\sum_{x,z} \chi\left(x + z + yx^d + yz^d\right) \right) \\
&= \sum_{x,z} \chi\left(x + z\right) \sum_y \chi\left(y\left(x^d + z^d\right)\right) \\
&= 2^n \cdot \sum_{x^d = z^d} \chi\left(x + z\right) \\
&= 2^n \cdot \left(1 + \sum_{x \neq 0} \sum_{\alpha^d = 1} \chi\left(x + \alpha x\right) \right) \\
&= 2^n \cdot \left(1 - r + \sum_x \sum_{\alpha^d = 1} \chi\left((1 + \alpha)x\right) \right) \\
&= 2^n \cdot \left(1 - r + 2^n\right),
\end{aligned}$$

which is what we wanted to show. □

Theorem 3.1 together with the details given by Cusick now implies the following theorem, which is stated as Conjecture 2 in [2].

THEOREM 3.3. *For odd k we have*

$$V(f_{-1}) = \frac{1}{3}(2^{2k+1} + 2^k - 1).$$

This theorem is closely related to Conjecture 4 of [2], which is still an open problem.

REFERENCES

[1] P. CHARPIN, *Cyclic codes with few weights and Niho exponents*, J. Combin. Theory Ser. A, 108 (2004), pp. 247–259.

[2] T. W. CUSICK, *Value sets of some polynomials over finite fields* GF(2^2m), SIAM J. Comput., 27 (1998), pp. 120–131.

[3] T. W. CUSICK AND P. MÜLLER, *Wan's bound for value sets of polynomials*, in Finite Fields and Applications (Glasgow, 1995), London Math. Soc. Lecture Note Ser. 233, Cambridge University Press, Cambridge, UK, 1996, pp. 69–72.

[4] H. DOBBERTIN, P. FELKE, T. HELLESETH, AND P. ROSENDAHL, *Niho type cross-correlation functions via Dickson polynomials and Kloosterman sums*, IEEE Trans. Inform. Theory, 52 (2006), pp. 613–627.

[5] H. DOBBERTIN, G. LEANDER, A. CANTEAUT, C. CARLET, P. FELKE, AND P. GABORIT, *Construction of bent functions via Niho power functions*, J. Combin. Theory Ser. A, 113 (2006), pp. 779–798.

[6] T. HELLESETH, *Some results about the cross-correlation function between two maximal linear sequences*, Discrete Math., 16 (1976), pp. 209–232.

[7] T. HELLESETH, J. LAHTONEN, AND P. ROSENDAHL, *On Niho type cross-correlation functions of m-sequences*, Finite Fields Appl., 13 (2007), pp. 305–317.

[8] T. HELLESETH AND P. ROSENDAHL, *New pairs of m-sequences with 4-level cross-correlation*, Finite Fields Appl., 11 (2005), pp. 674–683.

[9] Y. NIHO, *Multivalued Cross-Correlation Functions Between Two Maximal Linear Recursive Sequences*, Ph.D. thesis, University or Southern California, Los Angeles, CA, 1972.

[10] P. Rosendahl, *Niho Type Cross-Correlation Functions and Related Equations*, Ph.D. thesis, University of Turku, 2004, Turku, Finland; available online at http://www.tucs.fi/.
[11] P. Rosendahl, *A generalization of Niho's theorem*, Des. Codes Cryptogr., 38 (2006), pp. 331–336.
[12] D. Q. Wan, *A p-adic lifting lemma and its applications to permutation polynomials*, in Finite Fields, Coding Theory, and Advances in Communications and Computing (Las Vegas, NV, 1991), Lecture Notes in Pure and Appl. Math. 141, Dekker, New York, 1993, pp. 209–216.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 344–348

MATCHINGS AND NONRAINBOW COLORINGS*

ZDENĚK DVOŘÁK†, STANISLAV JENDROL'‡, DANIEL KRÁL'†, AND GYULA PAP§

Abstract. We show that the maximum number of colors that can be used in a vertex coloring of a cubic 3-connected plane graph G that avoids a face with vertices of mutually distinct colors (a rainbow face) is equal to $\frac{n}{2}+\mu^*-2$, where n is the number of vertices of G and μ^* is the size of the maximum matching of the dual graph G^*.

Key words. plane graphs, face-constrained coloring, nonrainbow coloring

AMS subject classification. 05C15

DOI. 10.1137/060675927

1. Introduction.

Colorings of embedded graphs with face-constraints have recently drawn the attention of several groups of researchers. The very first question that comes to one's mind in this area is the following.

QUESTION 1. *What is the minimal number of colors needed to color an embedded graph in such a way that each of its faces is incident with vertices of at least two different colors; i.e., there is no monochromatic face?*

This problem can be found in the work of Zykov [23] who studied the notion of *planar hypergraphs* and was further explored by Kündgen and Ramamurthi [16] for hypergraphs arising from graphs embedded in surfaces of higher genera. As an example of results obtained in this area, let us mention that every graph embedded on a surface of genus ε has a coloring with $O(\sqrt[3]{\varepsilon})$ colors [5] that avoids a monochromatic face.

An opposite type of question, motivated by results of anti-Ramsey theory, is the following.

QUESTION 2. *What is the maximal number $\chi_f(G)$ of colors that can be used in a coloring of an embedded graph G with no rainbow face; i.e., a face with vertices of mutually distinct colors?*

In our further considerations, we call a vertex coloring of G with no rainbow face a *nonrainbow coloring* of G. Notice that, unlike in the case of ordinary colorings, the goal in this scenario is to *maximize* the number of used colors. Though it may take some time to digest the concept, the setting is so natural that it has recently appeared independently in papers of Ramamurthi and West [20] and of Negami [17] (see also [1, 2, 18] for some even earlier results of this favor). In fact, Negami addressed the following extremal-type question (equivalent to Question 2).

*Received by the editors November 26, 2006; accepted for publication (in revised form) September 15, 2008; published electronically January 7, 2009.

http://www.siam.org/journals/sidma/23-1/67592.html

†Institute for Theoretical Computer Science (ITI), Faculty of Mathematics and Physics, Charles University, Malostranské náměstí 25, 118 00 Prague 1, Czech Republic (rakdver@kam.mff.cuni.cz, kral@kam.mff.cuni.cz). The Institute for Theoretical Computer Science is supported as project 1M0545 by Czech Ministry of Education.

‡Faculty of Science, Pavol Jozef Šafárik University in Košice, Jesenná 5, 041 54 Košice, Slovakia (stanislav.jendrol@upjs.sk). This author's work was supported by the Slovak Research and Development Agency under the contract APVV-0007-07.

§MTA-ELTE Egerváry Research Group (EGRES), Dept. of Operations Research, Eötvös University, Pázmány P. s. 1/C, Budapest, Hungary H-1117 (gyuszko@cs.elte.hu). This author's research was supported by OTKA grant K60802.

QUESTION 3. *What is the smallest number $k(G)$ of colors such that every vertex-coloring of an embedded graph G with $k(G)$ colors contains a rainbow face?*

It is not hard to see that $\chi_f(G) = k(G) - 1$ and the results obtained in either of the scenarios translate smoothly to the other one.

We now briefly survey results obtained in the direction of Questions 2 and 3 for planar graphs. Ramamurthi and West [21] noticed that every plane graph G has a nonrainbow coloring with at least $\alpha(G) + 1$ colors where $\alpha(G)$ is the independence number (stability) of G. In particular, every plane graph G of order n has a coloring with at least $\lceil \frac{n}{4} \rceil + 1$ colors by the Four Color Theorem. Also, Grötzsch's theorem [9, 22] implies that every triangle-free plane graph has a nonrainbow coloring with $\lceil \frac{n}{3} \rceil + 1$ colors. It was conjectured in [21] that this bound can be improved to $\lceil \frac{n}{2} \rceil + 1$. Partial results on this conjecture were obtained in [14] and the conjecture has eventually been proven in [12]. More generally, Jungić, Král', and Škrekovski [12] proved that every planar graph of order n with girth $g \geq 5$ has a nonrainbow coloring with at least $\lceil \frac{g-3}{g-2}n - \frac{g-7}{2(g-2)} \rceil$ colors if g is odd, and $\lceil \frac{g-3}{g-2}n - \frac{g-6}{2(g-2)} \rceil$ colors if g is even. All of these bounds are the best possible.

Complementary to the lower bounds on $\chi_f(G)$ presented in the previous paragraph, there are also results on upper bounds on $\chi_f(G)$. Negami [17] investigated nonrainbow colorings of plane triangulations G and showed that $\alpha(G) + 1 \leq \chi_f(G) \leq 2\alpha(G)$. In [6], it was shown that $\chi_f(G) \leq \lfloor \frac{7n-8}{9} \rfloor$ for n-vertex 3-connected plane graphs G, $\chi_f(G) \leq \lfloor \frac{5n-6}{8} \rfloor$ if $n \not\equiv 3 \pmod 8$, and $\chi_f(G) \leq \lfloor \frac{5n-6}{8} \rfloor - 1$ if $n \equiv 3 \pmod 8$ for 4-connected plane graphs G, and $\chi_f(G) \leq \lfloor \frac{43}{100}n - \frac{19}{25} \rfloor$ for 5-connected plane graphs G. The bounds for 3- and 4-connected graphs are the best possible.

Besides results on nonrainbow colorings of graphs with no short cycles and nontrivially connected plane graphs, there are also results on specific families on plane graphs, e.g., the numbers $\chi_f(G)$ were also determined for all semiregular polyhedra [11].

Let us mention that there are also results on mixed types of colorings in which we require that there is neither a monochromatic nor a rainbow face, e.g., [4, 13, 15]. For instance, it is known that each plane graph with at least five vertices has a coloring with two colors as well as a coloring with three colors that avoid both monochromatic and rainbow faces [3, 19].

The quantity $\chi_f(G)$ is also related to several parameters of the dual graph of G. In particular, $\frac{n}{2} + \mu^* - 2 \leq \chi_f(G) \leq n - \alpha^*$ for connected cubic plane graphs G [10], where α^* is the independence number of the dual graph G^* of G and μ^* is the size of the largest matching of G^*. In fact, it was conjectured that the first inequality is always an equality if G is 3-connected.

CONJECTURE 1. *The maximum number of colors used in a nonrainbow coloring of a cubic 3-connected plane graph G is related to the size of a maximum matching of its dual as follows:*

$$\chi_f(G) = \frac{n}{2} + \mu^* - 2.$$

We prove this conjecture. In our view, the fact that $\chi_f(G)$ only depends on the size of the largest matching of G^* in this specific case is quite surprising and deserves further investigation in a more general setting.

At the end of this paper, we briefly discuss generalizations and extensions of our results to cubic plane graph that need not be 3-connected. In particular, we show that the assumption that G is 3-connected cannot be relaxed.

2. Proof. If G is a plane graph, then $G^* = (V^*, E^*)$ denotes its plane dual and ϱ^* denotes the minimum size of an edge cover in G^*, i.e., the minimum size of a set of edges such that each vertex is incident with an edge in the set. Gallai's theorem relates the size of a maximum matching and the minimum edge-cover.

THEOREM 1 (see Gallai [7, 8]). *Let H be a graph without isolated vertices, μ the size of the maximum matching of H, and ρ the size of the minimum edge cover. The sum $\mu + \rho$ is equal to the number of vertices of H.*

By Euler's formula and Theorem 1, it holds that $\frac{n}{2} + \mu^* - 2 = n - \varrho^*$ for a 3-regular planar graph. Thus we prove the following theorem equivalent to the statement asserted in Conjecture 1.

THEOREM 2. *The maximum number of colors in a nonrainbow coloring of a cubic 3-connected planar graph $G = (V, E)$ is equal to $n - \varrho^*$.*

Proof. The easy part is to see that there is a nonrainbow coloring of that many colors. Let $E_C \subseteq E$ be the set of edges that corresponds to a minimum edge cover in G^*. The coloring is defined such that two vertices in V receive the same color if and only if they are in the same component of (V, E_c).

To prove the converse, we will rely on the min-max formula for edge cover, saying that the minimum size of an edge cover in a graph $G' = (V', E')$ without isolated vertices is equal to the maximum of $\sum_i \lceil \frac{1}{2}|K_i| \rceil$, where the maximum is taken over a vertex set $K \subseteq V'$, and K_i denotes the vertex sets of the components of $G'[K]$. This, for G^*, implies that

$$\varrho^* = \sum_i \left\lceil \frac{1}{2}|F_i| \right\rceil, \tag{1}$$

where, for some $F \subseteq V^*$, F_i are the components of $G^*[F]$. Let $V(F_i)$ denote the union of the boundaries of the faces in F_i, which is a subset of V. The sets $V(F_i)$ are disjoint. Hence, it suffices to prove for every nonrainbow coloring and every i that the number of colors appearing in $V(F_i)$ is no more than $|V(F_i)| - \lceil \frac{1}{2}|F_i| \rceil$. Fix an index i. Let $A_1, A_2, \ldots, A_k \subseteq V(F_i)$ denote the color-classes appearing in $V(F_i)$. We will prove that $k \leq |V(F_i)| - \lceil \frac{1}{2}|F_i| \rceil$, which thus concludes our proof.

We say that a color-class A_j claims a face, if the boundary of that face contains at least two vertices in A_j. Let $Z_j \subseteq F_i$ denote the set of faces in F_i that are claimed by A_j. Now consider the graph $H = (A_j, Q_j)$, where $Q_j = \{a_f b_f : f \in Z_j\}$, where a_f, b_f are two distinct vertices in $A_j \cap f$ (for every face $f \in Z_j$ choose one such pair of vertices). Hence, G is cubic, implying that H is subcubic. Thus

$$|Q_j| \leq \left\lfloor \frac{3}{2}|A_j| \right\rfloor. \tag{2}$$

Moreover, if $|A_j| = 1$, then $|Q_j| = 0$, and if $|A_j| = 2$, then $|Q_j| \leq 2$ (since G is 3-connected). By considering these inequalities, and inequality (2) in case of $|A_j| \geq 3$, we get that $|Z_j| = |Q_j| \leq 2(|A_j| - 1)$; i.e.,

$$|A_j| - 1 \geq \left\lceil \frac{1}{2}|Z_j| \right\rceil. \tag{3}$$

Thus

$$k = |V(F_i)| - \sum(|A_j| - 1) \leq |V(F_i)| - \sum \left\lceil \frac{1}{2}|Z_j| \right\rceil \leq |V(F_i)| - \left\lceil \frac{1}{2}|F_i| \right\rceil, \tag{4}$$

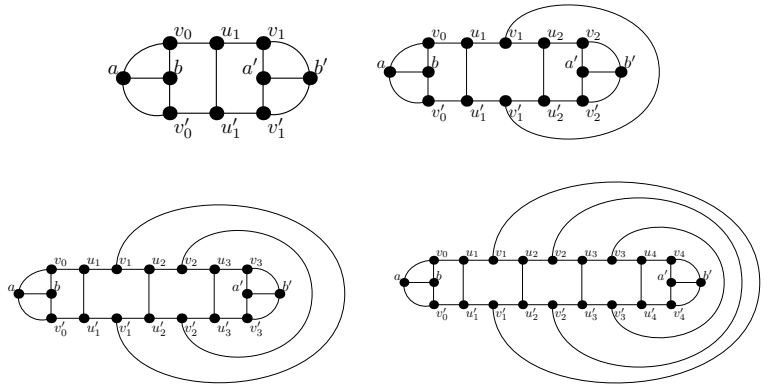


FIG. 1. *The graphs G_1, G_2, G_3, and G_4.*

where the last inequality follows from $|F_i| \leq \sum |Z_j|$ and the fact that the function $f(x) = \lceil \frac{x}{2} \rceil$ is superadditive over the natural numbers. ∎

Note that, as we discuss in section 4, the quantity $\chi_f(G)$ can be computed in polynomial time for 3-connected cubic plane graphs.

3. 2-connected cubic graphs. We first describe a construction of cubic 2-connected plane graphs G_ℓ, $\ell \geq 0$, which are our example graphs, given with a specific embedding into the plane. Start with two paths $v_0u_1v_1u_2v_2 \dots u_\ell v_\ell$ and $v'_0u'_1v'_1u'_2v'_2 \dots u'_\ell v'_\ell$ and add the edges $u_iu'_i$ for $i = 1, \dots, \ell$. Next, add an edge ab and join both a and b to both v_0 and v'_0. At the other ends of the paths, add an edge $a'b'$ and join both a' and b' to v_ℓ and v'_ℓ. Finally, add the edges $v_iv'_i$, $i = 1, \dots, \ell-1$, in such a way that they are drawn in the outer face. The graphs G_1, G_2, G_3, and G_4 can be found in Figure 1. Observe that the graph G_ℓ is a 2-connected cubic graph with $n = 4\ell + 6$ vertices and $f = 2\ell + 5$ faces. Also observe that the maximum matching of the dual graph G^*_ℓ has size $\ell + 2$.

THEOREM 3. *For every integer m, there exists a 2-connected cubic plane graph with n vertices such that*

$$\chi_f(G) > \frac{n}{2} + \mu^* - 2 + m.$$

Proof. Consider a graph G_ℓ for $\ell = 3m+1$ and color the following pairs of vertices with the same color (distinct pairs with distinct colors): a and b; a' and b'; u_i and u'_i for $i = 1, 4, \dots, \ell$; and v_i and v'_i for $i = 2, 5, \dots, \ell - 2$. Each of the remaining vertices gets a unique color. This way we construct a nonrainbow coloring of G with $4\ell + 6 - (2m + 3) = 10m + 7$ colors. Hence, we have the following:

$$\chi_f(G) - \frac{n}{2} - \mu^* + 2 \geq 10m + 7 - (6m + 5) - (3m + 3) + 2 = m + 1.$$

The statement of the lemma now follows. ∎

4. Concluding remarks. The statement of Theorem 2 cannot be extended to all cubic plane graphs without any further assumptions. Since the size of the maximum matching in a graph can be computed in polynomial time, it is possible to determine $\chi_f(G)$ for 3-connected cubic plane graphs in polynomial time. For bridgeless cubic plane graphs which need not be 3-connected, it seems natural to consider a dynamic programming approach based on the structure of cuts of sizes one and two in the graph G. Such an algorithm can utilize the following generalization of Theorem 2.

THEOREM 4. *If G is a plane 3-connected cubic graph and F a subset of its faces, then*

$$\chi_f^F(G) = n + \mu^* - |F|,$$

where $\chi_f^F(G)$ is the maximum number of colors that can be used in a coloring such that no face of F is rainbow, and μ^* is the size of a maximum matching of $G^*[F]$.

Though we believed that this approach should have led to a polynomial-time algorithm for determining $\chi_f(G)$ of all cubic graphs, we were not able to obtain such an algorithm; we suspect the problem could be NP-complete.

REFERENCES

[1] J. L. AROCHA, J. BRACHO, AND V. NEUMANN-LARA, *On the minimum size of tight hypergraphs*, J. Graph Theory, 16 (1992), pp. 319–326.

[2] J. L. AROCHA, J. BRACHO, AND V. NEUMANN-LARA, *Tight and untight triangulated surfaces*, J. Combin. Theory Ser. B, 63 (1995), pp. 185–199.

[3] A. A. DIWAN, *Disconnected 2-factors in planar cubic bridgeless graphs*, J. Combin. Theory Ser. B, 84 (2002), pp. 249–259.

[4] Z. DVOŘÁK AND D. KRÁL', *On planar mixed hypergraphs*, Electron. J. Combin., 8 (2001) p. R35.

[5] Z. DVOŘÁK, D. KRÁL', AND R. ŠKREKOVSKI, *Coloring face hypergraphs on surfaces*, European J. Combin., 26 (2005), pp. 95–110.

[6] Z. DVOŘÁK, D. KRÁL', AND R. ŠKREKOVSKI, *Non-rainbow colorings of 3-, 4-, and 5-connected plane graphs*, submitted.

[7] T. GALLAI, *Maximum-minimum Sätze über graphen*, Acta Math. Acad. Sci. Hungar., 9 (1958), pp. 395–434.

[8] T. GALLAI, *Über extreme Punkt- und Kantenmengen*, Ann. Univ. Sci. Budapest. Eötvös Sect. Math., 2 (1959), pp. 133–138.

[9] H. GRÖTZSCH, *Ein Dreifarbensatz für dreikreisfreie Netze auf der Kugel*, Wiss. Z. Martin-Luther-Universität, Halle, Wittenberg, Math.-Nat. Reihe, 8 (1959), pp. 109–120.

[10] S. JENDROL', *Rainbowness of cubic polyhedral graphs*, Discrete Math., 306 (2006), pp. 3321–3326.

[11] S. JENDROL' AND Š. SCHRÖTTER, *On rainbowness of semiregular polyhedra*, Czechoslovak Math. J., 58 (2008), pp. 359–380.

[12] V. JUNGIĆ, D. KRÁL', AND R. ŠKREKOVSKI, *Coloring of plane graphs with no rainbow faces*, Combinatorica, 26 (2006), pp. 169–182.

[13] D. KOBLER AND A. KÜNDGEN, *Gaps in the chromatic spectrum of face-constrained plane graphs*, Electron. J. Combin., 8 (2001), p. N3.

[14] D. KRÁL', *On maximum face-constrained coloring of plane graphs with no short face cycles*, Discrete Math., 277 (2004), pp. 301–307.

[15] A. KÜNDGEN, E. MENDELSOHN, AND V. VOLOSHIN, *Colouring planar mixed hypergraphs*, Electron. J. Combin., 7 (2000), p. R60.

[16] A. KÜNDGEN AND R. RAMAMURTHI, *Coloring face-hypergraphs of graphs on surfaces*, J. Combin. Theory Ser. B, 85 (2002), pp. 307–337.

[17] S. NEGAMI, *Looseness ranges of triangulations on closed surfaces*, Discrete Math., 303 (2005), pp. 167–174.

[18] S. NEGAMI AND T. MIDORIKAWA, *Loosely-tightness of triangulations of closed surfaces*, Sci. Rep. Yokohama Nat. Univ., Sect. I, Math. Phys. Chem., 43 (1996), pp. 25–41.

[19] J. G. PENAUD, *Une propriété de bicoloration des hypergraphes planaires*, in Colloque sur la Théorie des Graphes, Cahiers Centre Études Recherche Opér., 17 (1975), pp. 345–349.

[20] R. RAMAMURTHI AND D. B. WEST, *Maximum Face-Constrained Coloring of Plane Graphs*, Electronic Notes Discrete Math. 11, Elsevier, Amsterdam, 2002.

[21] R. RAMAMURTHI AND D. B. WEST, *Maximum face-constrained coloring of plane graphs*, Discrete Math., 274 (2004), pp. 233–240.

[22] C. THOMASSEN, *Grötzsch's 3-color theorem*, J. Combin. Theory Ser. B, 62 (1994), pp. 268–279.

[23] A. A. ZYKOV, *Hypergraphs*, Uspekhi Mat. Nauk, 29 (1974), pp. 89–154.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 349–368

GRAPH SEARCHING IN A CRIME WAVE*

DAVID RICHERBY† AND DIMITRIOS M. THILIKOS‡

Abstract. We define helicopter cops and robber games with multiple robbers, extending previous research, which considered only the pursuit of a single robber. Our model is defined for robbers that are visible (their position in the graph is known to the cops) and active (they can move at any point in the game) but is easily adapted to other variants of the single-robber game that have been considered in the literature. We show that the game with many robbers is nonmonotone: that is, fewer cops are needed if the robbers are allowed to reoccupy positions that were previously unavailable to them. As the moves of the cops depend on the position of the visible robbers, strategies for such games should be interactive, but the game becomes, in a sense, less interactive as the initial number of robbers increases. We prove that the main parameter emerging from the game, which we denote $\mathbf{mvams}(G, r)$, captures a hierarchy of parameters between proper pathwidth and proper treewidth, and we completely characterize it for trees, extending analogous existing characterizations of the pathwidth of trees. Moreover, we prove an upper bound for $\mathbf{mvams}(G, r)$ on general graphs and show that this bound is reached by an infinite class of graphs. On the other hand, if we consider the robbers to be invisible and lazy, the resulting parameters collapse in all cases to either proper pathwidth or proper treewidth, giving a further case where the classical equivalence between visible, active robbers and invisible, lazy robbers does not hold.

Key words. graph searching, treewidth, pathwidth

AMS subject classifications. 05C83, 05C85

DOI. 10.1137/070705398

1. Introduction.

During recent decades, the problem of searching a graph has attracted much attention not only because of its purely graph-theoretic interest but also for its numerous applications in modeling problems in communication networks (for related surveys, see [1, 2]). In general, graph searching problems are described in terms of a game played between a team of cops and a robber, whom the cops attempt to capture by moving systematically through the graph. We wish to know the minimum number of cops required to catch the robber, subject to various constraints on their behavior and that of the robber. Several versions of the game have been examined, differing, for example, in whether the cops know the position of the robber, whether the robber can move at will or only when disturbed by a cop, and how the cops can move through the graph.

One of the main models of graph searching, known as the helicopter cops and robber game, was introduced by Seymour and Thomas [11]. In this model, the robber occupies a vertex of a graph and is *active* in the sense that he may move at every round of the game along any path of any length whose vertices are not guarded by the cops. On the other hand, the cops are not constrained to stay within the graph and can be placed on or removed from vertices of the graph, as if flying by helicopter.

*Received by the editors October 15, 2007; accepted for publication (in revised form) September 29, 2008; published electronically January 7, 2009. This research carried out while the first author was a post-doc of the Graduate Program in Logic, Algorithms, and Computation ($\mu\prod\lambda\forall$) at the Department of Mathematics, National and Kapodistrian University of Athens.

http://www.siam.org/journals/sidma/23-1/70539.html

†School of Computing, University of Leeds, Leeds, LS2 9JT, UK (richerby@comp.leeds.ac.uk). This author was funded by the European Social Fund and Greek National Resources (ΕΠΕΑΕΚ ΙΙ) PYTHAGORAS II.

‡Department of Mathematics, National and Kapodistrian University of Athens, Panepistimioupolis, GR-15784 Athens, Greece (sedthilk@math.uoa.gr).

A crucial feature of this game is that the robber is *visible*: the cops have complete knowledge of his current position. Victory for the cops is declared when a cop lands on the vertex occupied by the robber and the robber cannot make any move to escape. Since the cops base their moves on the current position of the robber, the strategy they use is *interactive*. In [11], Seymour and Thomas proved that the minimum number of cops guaranteed to be able to win the game is one greater than the treewidth of the graph on which it is played. The proof of this result includes a proof of the monotonicity of the game, i.e., that the cops do not become weaker when their moves are restricted to those that monotonically decrease the portion of the graph available to the robber.

Variants of the above game were considered in [6], where now the robber is a *lazy* fugitive who moves only when a cop lands on the vertex he occupies. However, to compensate, the robber is now *invisible*: his position is unknown to the cops. Notice that in this game, the cops' strategy is predetermined and can be given in advance. Games defined with this characteristic are described as *fugitive games* in order to stress the invisibility of the robber. This second version is equivalent to the Seymour–Thomas game in the sense that, for any graph, the two games require the same number of cops [6]. It follows easily from [6] (see also [3]) that, when the fugitive is active and invisible, the number of cops required to ensure his capture is one greater than the pathwidth of the graph—another graph parameter of equal importance to treewidth.

This paper intends to examine, and also unify, the above models under the natural extension where the graph contains many robbers rather than just one. This is the first time that multiple robbers have been considered in graph searching, and we believe that our results will motivate such a study for other models as well.

We describe our model for graph searching using the most general setting of *mixed searching*, proposed by Bienstock and Seymour [3] and also examined in [12, 13, 14, 15]. In this model, each move of the cops consists either of a placement or removal (as before) or of *sliding* a cop along an edge of the graph. This may reduce by one the number of cops required to search a graph, but, as observed in [3], the version without sliding can be reduced to a mixed search by replacing each edge in the graph with two parallel edges (or a triangle involving a new vertex). Moreover, apart from being more general, including sliding in our model makes the presentation of our results cleaner.

It is not obvious how to generalize the concept of monotonicity to the setting with many robbers. Now, each robber has his own individual free space, leading to the question of whether monotonicity should be defined individually or collectively. We give three natural definitions and show them to be equivalent.

Monotonicity is crucial in the multiple-robber case. If we do not require monotonicity, we can catch any number r of visible, active robbers one at a time by repeating the strategy to catch a single robber, without requiring any additional cops. However, when we restrict our attention to monotone strategies, the number of cops required, which we denote $\mathbf{mvams}(G, r)$ (for monotone, visible, active, mixed search number against r robbers), can be greater than the nonmonotone case and depends on the number of robbers. In particular, $\mathbf{mvams}(G, 1)$ is just the mixed search number for a single visible active robber. This, in turn, is equal to the parameter of *proper treewidth* defined in [5,12]. On the other hand, if n is the number of vertices in G, then $\mathbf{mvams}(G, n)$ is equal to the mixed search number for a single invisible active robber which, in turn, corresponds to the parameter of *proper pathwidth* defined in [14]. Moreover, we show that $\mathbf{mvams}(G, r)$ can, for appropriate values of r, take all intermediate values between proper treewidth and proper pathwidth. As our main result, we give the exact value of $\mathbf{mvams}(G, r)$ on trees and an upper bound for general

graphs:

$$\mathbf{mvams}(T,r) = \min\{\mathbf{ppw}(T), \lfloor \log r \rfloor + 1\} \qquad \text{(for any tree } T\text{)},$$
$$\mathbf{mvams}(G,r) \le \min\{\mathbf{ppw}(G), \mathbf{ptw}(G)\cdot(\lfloor \log r \rfloor + 1)\} \qquad \text{(for any graph } G\text{)},$$

where $\mathbf{ppw}(G)$ and $\mathbf{ptw}(G)$ denote, respectively, the proper pathwidth and proper treewidth of the graph G.

Our result for trees is based on a complete characterization of $\mathbf{mvams}(T,r)$ on trees and extends the analogous characterizations for pathwidth and proper pathwidth given in [13] and [7], respectively.

Our results can be seen as showing that the number of robbers tunes the amount of interactivity in search strategies, spanning all intermediate levels from pathwidth (fully predetermined) to treewidth (fully interactive). A rather different way of defining this tuning was given by Fomin, Fraigniaud, and Nisse, who considered a single active robber but restricted the number of rounds at which the cops can ask for the robber's position [8].

A natural question is whether the same variation of values can be achieved in the setting of invisible but lazy fugitives defined in [6], given that a single invisible lazy fugitive is equivalent to a single visible active robber. However, in the case of multiple robbers, being lazy and invisible is not the same as being active and invisible. Here, we can define laziness as meaning either that a robber may move only when the cops land on his vertex or that all robbers may move together when a cop lands on any single robber. For either definition of laziness, with or without monotonicity, the hierarchy collapses, and, in a graph of order n, any number of robbers is equivalent to either a single robber or n robbers. Thus, the multiple-robber setting is degenerate for games with predetermined strategies, which supports our decision to consider the interactive strategies generated by the visible, active setting. Note that this is not the first case where invisibility cannot be exchanged for laziness: Hunter and Kreutzer have shown that the symmetry breaks, even for one robber, when the games are defined on directed graphs [10].

The remainder of the present paper is organized as follows. Our graph searching model is defined in detail in section 2. In section 3, we show the equivalence of three reasonable definitions of monotonicity and explore the role of monotonicity in the game. To relate our hierarchy of parameters to the well-known parameters of proper pathwidth and proper treewidth, we make a brief detour through the theory of games with an invisible robber in section 4, where we also show that the case of multiple invisible robbers collapses to already-studied cases. In section 5, we give upper bounds for the number of cops required to catch r robbers in trees and in general graphs, and, in section 6, we show that the upper bound for trees is, in fact, an exact characterization of the number of cops needed. We also show that the upper bound for general graphs is reached by an infinite class of graphs. Several consequences and open problems emerging from our results are presented in section 7.

2. The searching model. All graphs considered in this paper are finite, simple, and, unless otherwise stated, undirected.

In a helicopter search game with many visible robbers, the opponents are a group of k cops and a group of r robbers, who occupy vertices of the graph. The goal of the cops is to capture all of the robbers. At all times, the cops and robbers have full information about each other's location and may use this information to decide their next move. Initially, there are no cops in the graph, but, at all times, any robber who has not been captured is on some vertex.

A play of the game consists of a sequence of rounds, with each round consisting of three parts, as follows.

Announcement. The cops announce their intended move to the robbers. One cop moves in each round, by one of the following operations.

- Placement of a cop on a vertex v, not currently occupied by a cop. The move is denoted by $\mathbf{place}(v)$.
- Removal of a cop from an occupied vertex v, denoted by $\mathbf{remove}(v)$.
- Sliding of a cop from the one endpoint u of an edge $\{u, v\}$ to the other, which is initially not occupied by a cop. The move is denoted by $\mathbf{slide}(u \to v)$.

Avoidance. Each robber who has not yet been captured can move with infinite speed to any vertex reachable from his current position by a path not blocked by cops, as long as this vertex will not be occupied by a cop once the cops' current move has been realized. The robbers are "active" in the sense that any robber may move in the graph at any move of the game, as long as he has an unblocked path to move along.

If the announced move is a placement to or removal from some vertex, that vertex is not considered to be blocked for the purposes of the robbers' movement in the round. If the announced move is $\mathbf{slide}(u \to v)$, the edge uv is considered to be blocked for this round but the vertices u and v are not.

Realization. The cops carry out the announced action.

A robber is captured if the cops announce that they will move (by placement or sliding) to the vertex he occupies and there is no way for him to move to another vertex.

To formalize the game, we will use a string $\mathbf{R} \in (V(G) \cup \{*\})^r$ to denote the positions of the r robbers in the graph. In particular, the ith character of $\mathbf{R}$ is either the vertex occupied by the ith robber or "$*$" in the case that the ith robber has been captured. We write $V(\mathbf{R})$ for the set of characters in $\mathbf{R}$, other than $*$. Since, at any time, there is at most one cop on any vertex, we may represent the position of the cops as a set $S \in V(G)^{[\leq k]}$.

A *play* of the game on a graph is an infinite sequence of positions

$$\mathcal{P} = S_0, \mathbf{R}_0, S_1, \mathbf{R}_1, \dots,$$

where, for each i, the transition from having the cops at S_i and robbers at $\mathbf{R}_i$ to the cops at S_{i+1} and robbers at $\mathbf{R}_{i+1}$ is a valid move of the game, as described above. Specifically, the sequence $S_0, S_1, \dots$ has the properties that

- $S_0 = \emptyset$;
- $S_1 = \{v\}$ for some vertex v—the first move is $\mathbf{place}(v)$; and
- for consecutive sets S_i and S_{i+1}, one of the following holds:
  - $S_{i+1} - S_i = \{v\}$—the move is $\mathbf{place}(v)$,
  - $S_i - S_{i+1} = \{v\}$—the move is $\mathbf{remove}(v)$,
  - $S_{i+1} \triangle S_i = \{u, v\} \in E(G)$—the move is $\mathbf{slide}(u \to v)$, where $S_i - S_{i+1} = \{u\}$ and $S_{i+1} - S_i = \{v\}$.

We call such a sequence of cop positions *consistent*.

Given two consecutive sets S_i and S_{i+1} of a consistent sequence, we say that a path P of G is (S_i, S_{i+1})-*avoiding* if its internal vertices avoid $S_i \cap S_{i+1}$, its last vertex is not in S_{i+1}, and, in the case that $|e| = 2$, its edges avoid the edge $e = S_{i+1} \triangle S_i$.

Given that the location of the robbers at the ith step is $\mathbf{R}_i = [a_1 \dots a_r]$, we define the set of free locations for the jth robber after step i as $F^j_{i+1} = \emptyset$ if $a_j = *$ and, otherwise,

$$F^j_{i+1} = \big\{ y \in V(G) - S_{i+1} \mid G \text{ contains an } (S_i, S_{i+1})\text{-avoiding } (a_j, y)\text{-path} \big\}.$$

As a response to the ith move of the cops, the robbers can choose their new location to be any string $\mathbf{R}_{i+1} = [a'_1 \dots a'_r]$ such that for $j \in \{1, \dots, r\}$, $a'_j = *$ if $F^j_{i+1} = \emptyset$ and $a'_j \in F^j_{i+1}$ otherwise. (In particular, note that, if $a_j = *$, then $a'_j = *$ also.)

We set $F_0 = V(G)$, and for $i \geq 1$, we define $F_i = \bigcup_{j \in \{1,\dots,r\}} F^j_i$. We say that the sequence $F_0, F_1, \dots$ is the sequence of *free positions* for the robbers. If, for every $i \geq 0$, $F_{i+1} \subseteq F_i$, we say that $\mathcal{P}$ is a *monotone* play. (Other definitions of monotonicity are considered in section 3 and shown to be equivalent to this definition in the sense that they lead to the same graph parameter.)

A play $\mathcal{P} = S_0, \mathbf{R}_0, S_1, \mathbf{R}_1, \dots$ is *winning* (for the cops) if $V(\mathbf{R}_i) = \emptyset$ for some $i \geq 0$; that is, all the robbers are eventually captured. The *essential part* of a winning play is the subsequence $S_0, \mathbf{R}_0, \dots, S_\ell, \mathbf{R}_\ell$, where ℓ is minimal such that $V(\mathbf{R}_\ell) = \emptyset$.

According to our description of the game, any move of the cops may depend on the current position of the cops and robbers in the graph. A *search strategy of cost k against r robbers* or, more succinctly, a *(k, r)-strategy* is a function

$$\mu : V(G)^{[\leq k]} \times (V(G) \cup \{*\})^r \to V(G)^{[\leq k]},$$

whose inputs are the position S of the cops and the positions $\mathbf{R}$ of the robbers and whose output is S', the new position of the cops, such that, for all S and $\mathbf{R}$, the sets S and S' obey the restrictions given in the definition of consistency for sequences. That is, there is a single move which transforms the cop position S to S'.

Note that, when we define strategies, we will not define the action of the cops in positions that can never occur when the strategy is executed. Thus, we give only a partial function. Formally, the strategy is any total extension of this partial function, assigning arbitrary moves to the cops in situations that do not occur in any play in which the cops follow the given partial strategy.

A *play with respect to a (k, r)-strategy μ*, or a *μ-play*, is any play $S_0, \mathbf{R}_0, S_1, \mathbf{R}_1, \dots$ where $S_{i+1} = \mu(S_i, \mathbf{R}_i)$ for all $i \geq 0$. A strategy μ is said to be *monotone* if all μ-plays are monotone and *winning* if all μ-plays are winning.

We define the *nonmonotone* and *monotone visible active mixed search number*, respectively, of a graph G as follows:

$$\mathbf{vams}(G, r) = \min \big\{ k \mid \text{there is a winning } (k, r)\text{-strategy on } G \big\},$$
$$\mathbf{mvams}(G, r) = \min \big\{ k \mid \text{there is a monotone winning } (k, r)\text{-strategy on } G \big\}.$$

To describe a search strategy as even a partial function is often rather cumbersome. Instead, we will frequently describe a search strategy as a *search program* Π that makes move decisions depending only on the current position of the cops and robbers, without reference to previous positions in the search. Thus, we can extract a strategy from a search program and vice versa. We call a search program monotone or winning if the corresponding search function is. The program receives the information on the positions of the robbers by calling a routine *robbers_positions()*.

As an example we give program 1, a monotone winning search program for one cop against one robber in a tree T. Notice that, at each step, the robber must choose

Search program 1. $\Pi(T, 1)$ to capture one robber in a tree T.

place(v) where v is any vertex of T.
Let $\mathbf{R} \leftarrow$ *robbers_positions()*.
Let $T' \leftarrow T$.
While $V(\mathbf{R}) \neq \emptyset$,
    Let T' be the connected component of $T - v$ containing $V(\mathbf{R})$
        and let w be the (unique) vertex of T' adjacent to v.
    slide$(v \to w)$.
    Let $v \leftarrow w$.
    Let $\mathbf{R} \leftarrow$ *robbers_positions()*.
remove(v).

his position in the connected component T' where he resides, excluding the vertex w that is the target of the cop's move. At each round, the set of free positions of the robber becomes strictly smaller, ensuring both monotonicity and the eventual capture of the robber.

We may also represent a winning (k, r)-strategy μ as a finite tree. Let T_μ be the least labeled, rooted, directed tree with the following properties. (By "least," we mean that no proper subtree of T_μ can be labeled to meet our requirements. Note that there may be more than one vertex or edge with any given label and that, when we speak of a path in T_μ, we mean a maximal directed path from the root to a leaf. We could also represent nonwinning strategies as infinite trees in a similar way, but we need only representations of winning strategies.)

- Every edge is directed away from the root.
- Every vertex is labeled with a set $S \in V(G)^{[\leq k]}$, and every edge is labeled with a string $\mathbf{R} \in (V(G) \cup \{*\})^r$.
- The essential part of every μ-play in G labels some path in T_μ.

Notice that, according to the above, the root of T_μ is labeled by the empty set, corresponding to the position of the cops at the beginning of any μ-play.

Our manipulation of tree representations of strategies will often lead us to construct trees that do not represent strategies because they are *defective* in some way. Allowing such trees makes several of our proofs more straightforward. Here, we describe the defects that may arise and show how to repair them.

Nondeterminism. A vertex v labeled S might have distinct outgoing edges vv' and vv'' with the same label $\mathbf{R}$ but with v' and v'' having labels S' and S'' (where S' and S'' are not necessarily distinct). Thus, with the cops in position S and the robbers at $\mathbf{R}$, the cops can win by moving to either S' or S''. Hence, we may delete the subtree rooted at v''.

Null moves. A vertex v labeled S might have a child v' also labeled S. This corresponds to the cops deciding to do nothing for a move. Since the robbers may move anywhere in their free space, allowing them to make two consecutive moves while the cops stay still gives them no extra power. Hence, we may delete the vertex v' and, for every child w of v', where the edge $v'w$ is labeled $\mathbf{R}$, add an edge vw, also labeled $\mathbf{R}$.

Inconsistency. There may be distinct vertices v and w, both labeled S, with outgoing edges vv' and ww', respectively, that have the same label $\mathbf{R}$ but with v' and w' having distinct labels S' and S'', respectively. As in the case of nondeterminism,

this means that the cops have a choice of ways to win from the position $(S, \mathbf{R})$. We may replace the subtree rooted at w' with a copy of the subtree rooted at v'.

Repeated application of these operations will yield a tree that properly corresponds to a winning strategy. Further, the resulting strategy uses no more cops than were deployed in the original defective tree and is monotone if and only if every play in the defective tree was monotone.

Note, in particular, that the discussion above of inconsistent trees justifies our decision to define strategies as functions:

$$\mu : V(G)^{[\leq k]} \times (V(G) \cup \{*\})^r \to V(G)^{[\leq k]}.$$

Such strategies are known as *positional* or *memoryless* strategies: they determine the move of the cops solely from the current position in the game. One could define a *general strategy* to be a function that chooses the moves based on the full history of the game, i.e., a function

$$M : \left(V(G)^{[\leq k]} \times (V(G) \cup \{*\})^r\right)^{[<\omega]} \to V(G)^{[\leq k]}.$$

The tree associated with such a strategy may be inconsistent: for example, the move made in a position with two cops on the graph may depend on which of the cops was last to move. However, given the tree associated with a general winning strategy M, we can produce a winning strategy μ that uses the same number of cops and that is monotone if M is. We summarize the above observations with the following.

PROPOSITION 1. *There is a winning (k, r)-strategy μ for a graph G if and only if there is a winning general (k, r)-strategy M for G. Further, μ may be chosen to be monotone if M is monotone.*

In program 1, the moves of the cops depend only on the knowledge of which component of the tree contains the robber and not on the precise vertex he occupies. With an eye to the situation with more than one robber, we can say that the move of the cops from position S depends only on the knowledge of how many robbers are in each component of $T - S$.

In fact, the cops do not lose any strength if their information is restricted in this way. For this, given an $S \in V(G)^{[\leq k]}$ and $\mathbf{R}, \mathbf{R}' \in (V(G) \cup \{*\})^r$, we say that $\mathbf{R} \equiv_S \mathbf{R}'$ if every component of $G - S$ that contains m robbers in $\mathbf{R}$ also contains m robbers in $\mathbf{R}'$ (where $G - S$ is the graph that results from deleting the vertices in S from G). Notice that $\equiv_S$ is an equivalence relation. We call a (k, r)-strategy *smooth* if, for every $\mathbf{R}, \mathbf{R}' \in (V(G) \cup \{*\})^r$ where $\mathbf{R} \equiv_S \mathbf{R}'$, we have $\mu(S, \mathbf{R}) = \mu(S, \mathbf{R}')$. That is, the cops' moves depend only on the number of robbers in each component of $G - S$ and not on their locations within these components.

LEMMA 2. *There is a winning (k, r)-strategy in G if and only if there is a smooth winning (k, r)-strategy in G.*

Proof. Let μ be a winning (k, r)-strategy in G. For each $S \in V(G)^{[\leq k]}$, let $\mathcal{A}_S$ be the set of $\equiv_S$-equivalence classes of robber positions. For each $A \in \mathcal{A}_S$ we select an arbitrary representative $\mathbf{R}_{A,S}$. Now, define a strategy μ' by putting $\mu'(S, \mathbf{R}) = \mu(S, \mathbf{R}_{A,S})$ whenever $\mathbf{R} \equiv_S \mathbf{R}_{A,S}$.

It is clear that μ' is smooth; it remains to show that it is winning. Let $\mathcal{P}' = S'_0, \mathbf{R}'_0, S'_1, \mathbf{R}'_1, \ldots$ be any μ'-play. From the definition of μ', there is a μ-play $\mathcal{P} = S'_0, \mathbf{R}_0, S'_1, \mathbf{R}_1, \ldots$ such that, for all $i \geq 0$, $\mathbf{R}_i \equiv_{S_i} \mathbf{R}'_i$. Since the possible moves of each robber depend on his free space and not on his precise position in the graph, any move by the cops that captures a robber in $\mathcal{P}$ must also capture a robber in $\mathcal{P}'$. $\mathcal{P}$ is a winning play, so $\mathcal{P}'$, and hence μ', must also be winning. ∎

Note that the strategies referred to when considering smoothness are not necessarily monotone but that, in the above proof, μ' is monotone if μ is.

Finally, in this section, we show that the parameters we have defined are closed under taking minors. Recall that G is a minor of H (written $G \preccurlyeq H$) if G can be constructed from H by a sequence of vertex deletions, edge deletions, and edge contractions, where an edge contraction is the deletion of two adjacent vertices u and v in H, followed by the addition of a new vertex w adjacent to all former neighbors of the deleted vertices.

PROPOSITION 3. *If $G \preccurlyeq H$, then, for any r,* $\mathbf{mvams}(G, r) \leq \mathbf{mvams}(H, r)$.

Proof. Let μ be a smooth, monotone, winning (k, r)-strategy for H, and let T_μ be the tree representing μ. We may assume that G is formed from H by deleting a single isolated vertex or deleting or contracting a single edge, since deletion of a nonisolated vertex may be achieved by first deleting all its edges.

Suppose $G = H - v$ for some $x \in V(H)$. Since x is isolated in H, any move of the cops involving x must be either a placement or a removal. Let T'_μ be the tree that results from deleting x from every vertex label in T_μ and replacing x with $*$ in every edge label. Clearly, T'_μ is a (possibly defective) tree corresponding to a monotone (k, r)-strategy for G.

Suppose $G = H - e$ for some edge $e = xy \in E(H)$. The only alterations we need to make to T_μ are to deal with slides along the now-deleted edge. Suppose $v \in V(T_\mu)$ is labeled S and sends an edge labeled $\mathbf{R}_1$ to vertex v_1, labeled S', such that $S \triangle S' = e$. Let $\mathbf{R}_1, \ldots, \mathbf{R}_\ell$ enumerate the $\equiv_S$-equivalence class of $\mathbf{R}_1$. By smoothness, v also has children $v_2, \ldots, v_\ell$ such that the edge vv_i is labeled $\mathbf{R}_i$ and v_i is labeled S' for each $i \in \{2, \ldots, \ell\}$. We may assume, without loss of generality, that $x \in S$, i.e., that the slide is from x to y. By monotonicity, the only neighbor of x in H that can be in the robbers' free space is y. Therefore, in G, no neighbor of x is in the robbers' free space, and we can replace the slide $x \to y$ with a removal from x followed by a placement to y. For each $i \in \{1, \ldots, \ell\}$, add a new vertex w_i, labeled $S - x$, and an edge vw_i, labeled $\mathbf{R}_i$. For each $j \in \{1, \ldots, \ell\}$, make a copy of the subtree of T_μ rooted at v_j, and add an edge labeled $\mathbf{R}_j$ from w_i to the root of the jth copy.

Finally, suppose G is the result of contracting the edge xy in H to give a new vertex which we denote v_{xy}. To construct a (possibly defective) strategy tree for G, it suffices to substitute v_{xy} for both x and y in all vertex and edge labels in T_μ. A robber on v_{xy} in G can reach any vertex reachable by a robber on x or y in H. The effect for the cops is as follows:

- $\mathbf{place}(x)$ becomes a null move if there was already a cop on y and, if not, becomes $\mathbf{place}(v_{xy})$;
- $\mathbf{slide}(x \to y)$ becomes a null move;
- for $z \neq y$, $\mathbf{slide}(x \to z)$ becomes $\mathbf{place}(z)$ if there is a cop on y and, if not, becomes $\mathbf{slide}(v_{xy} \to z)$;
- for $z \neq y$, $\mathbf{slide}(z \to x)$ becomes $\mathbf{remove}(z)$ if there is a cop on y and, if not, becomes $\mathbf{slide}(z \to v_{xy})$;
- $\mathbf{remove}(x)$ becomes a null move if there is a cop on y and, if not, becomes $\mathbf{remove}(v_{xy})$.

The cases for moves involving y are symmetric. ∎

3. Variants of monotonicity. In the previous section, we defined the concept of monotonicity for plays and strategies. These definitions are natural extensions of

the case with only one robber but are not the only ones. In this section, we consider two further natural definitions of monotonicity, which turn out to be equivalent to our first definition, and we begin an investigation of the cost of monotonicity.

Let $\mathcal{P} = S_0, \mathbf{R}_0, S_1, \mathbf{R}_1, \ldots$ be a μ-play. We say that $\mathcal{P}$ is *pointwise monotone* if, for each $j \in \{1, \ldots, r\}$ and each $i \geq 0$, $F^j_{i+1} \subseteq F^j_i$; i.e., no single robber's set of free positions ever increases. Also, we say that $\mathcal{P}$ is *cop-monotone* if, for each $v \in V(G)$, the set

$$s_{\mathcal{P}}(v) = \{i \mid v \in S_i \text{ and } V(\mathbf{R}_i) \neq \emptyset\}$$

is an interval of $\mathbb{N}$—that is, once the cops have left a vertex, they never return to it as long as there are robbers in the graph. Observe that any cop-monotone μ-play must be a winning μ-play because plays are infinite and G is not, so the cops must eventually revisit a vertex if the robbers live forever. We say that a (k, r)-strategy μ is monotone according to one of the above definitions if all μ-plays are.

LEMMA 4. *Let G be a graph, and let k and r be positive integers. The following are equivalent:*

1. *there is a monotone winning (k, r)-strategy in G;*
2. *there is a pointwise-monotone winning (k, r)-strategy in G;*
3. *there is a cop-monotone (k, r)-strategy in G.*

Proof. (2) $\Rightarrow$ (1) follows trivially from the definitions.

(3) $\Rightarrow$ (2). Let μ be a cop-monotone strategy, and suppose that, for some μ-play $\mathcal{P} = S_0, \mathbf{R}_0, S_1, \mathbf{R}_1, \ldots$, there is a step, say, from S_i to S_{i+1}, where the free space of the jth robber ($1 \leq j \leq r$) increases, i.e., that $F^j_{i+1} \supset F^j_i$. It follows that there must have been a cop removed or slid from some vertex $v \in \partial_G(F^j_i)$, the set of vertices in $V(G) - F^j_i$ that are adjacent to at least one vertex in F^j_i. There is a μ-play $\mathcal{P}'$ that follows $\mathcal{P}$ until S_{i+1} and in which the jth robber moves to the newly vacated vertex v and stays there at all subsequent moves. But now, this robber cannot be caught unless v is revisited, contradicting the assumed cop-monotonicity of μ.

For (1) $\Rightarrow$ (3), the idea is that the cops never need to visit a vertex that is not in the robbers' free space because such a move can never decrease the free space. Therefore, the move does not contribute to the capture of the robbers and can safely be omitted. Thus, every move made by the cops may be assumed to be a removal, a placement, or a slide into the robbers' free space and, since the free space is monotonically decreasing, each such move must be the first time the target vertex has been visited.

Formally, let μ be a monotone winning strategy which we may assume, by Lemma 2, to be smooth. Let $\mathcal{E}$ be the set of the essential parts of all μ-plays. For any μ-play $\mathcal{P} \in \mathcal{E}$, let $c(\mathcal{P})$ be the number of vertices revisited by the cops (i.e., the number of vertices v for which $s_{\mathcal{P}}(v)$ is not an interval), and let $c(\mu) = \sum_{\mathcal{P} \in \mathcal{E}} c(\mathcal{P})$. $c(\mu)$ is well defined, as $\mathcal{E}$ is finite; further, $c(\mu) = 0$ if and only if μ is cop-monotone.

Suppose μ is not cop-monotone. We construct a (k, r)-strategy μ' with $c(\mu') < c(\mu)$. Repeated applications of this transformation will yield the desired cop-monotone strategy.

Let $\mathcal{P} = S_0, \mathbf{R}_0, S_1, \mathbf{R}_1, \ldots$ be a non–cop-monotone μ-play, and let v_{i+1} be a vertex that is revisited for the first time in the step from S_i to S_{i+1}. This means that $S_{i+1} - S_i = \{v_{i+1}\}$—the move is a placement or a slide. Let H be the union of the connected components of $G - S_i$ that intersect $V(\mathbf{R}_i)$, and let $S^* = \partial_G H$. Notice that $S^* \subseteq S_i$ since, otherwise, $\mathcal{P}$ is not monotone.

We cannot have $v_{i+1} \in V(H)$ since, by monotonicity, none of the vertices in H has yet been visited. We cannot have $v_{i+1} \in S^*$ as there are already cops on every vertex of S^*, so these vertices cannot be the target of a placement or a slide. Further, if the move is a slide, from the single vertex $v_i \in S_i - S_{i+1}$, then $v_i \notin S^*$: suppose $v_i \in S^*$; since $v_{i+1} \notin H$, v_i becomes part of the robbers' free space, contradicting the monotonicity of μ.

Let $T_0 = S_i$ and, for $j \geq 0$, let $T_{j+1} = \mu(T_j, \mathbf{R}_i)$. Let h be minimal such that $T_{h+1} \cap V(H) \neq \emptyset$. Thus, $T_1 = S_{i+1}$ and T_h is the first move at which the cops will play into the robbers' free space if the robbers stand still. Notice that, in any μ-play that includes the position $S_i, \mathbf{R}_i$, the positions $T_1, \ldots, T_h$ will follow because, by smoothness, the moves of the cops depend only on the free space of the robbers, which does not change during the quoted sequence (indeed, this remains true if we replace $\mathbf{R}_i$ with any $\mathbf{R} \equiv_{S_i} \mathbf{R}_i$). As before, the monotonicity of μ implies that $S^* \subseteq T_j$ for $j \in \{0, \ldots, h\}$.

We now define μ'. The idea is to replace the sequence $T_0, \ldots, T_h$ with a new sequence of moves that performs the least number of removals and placements to move the cops from T_0 to T_h but omits the move to v_{i+1}. Since none of these moves is in H, the robbers' free space remains the same and monotonicity is preserved. Toward this end, let $T_0 - T_h = \{x_1, \ldots, x_p\}$ and $T_h - T_0 - \{v_{i+1}\} = \{y_1, \ldots, y_q\}$. For any $\mathbf{R} \equiv_{S_i} \mathbf{R}_i$, set

$$\begin{aligned} \mu'(T_0, \mathbf{R}) &= T_0 - \{x_1\}, \\ \mu'(T_0 - \{x_1\}, \mathbf{R}) &= T_0 - \{x_1, x_2\} \\ &\vdots \\ \mu'(T_0 - \{x_1, \ldots, x_{p-1}\}, \mathbf{R}) &= T_0 - \{x_1, \ldots, x_p\}. \end{aligned}$$

Writing T' for $T_0 - \{x_1, \ldots, x_p\}$, set

$$\begin{aligned} \mu'(T', \mathbf{R}) &= T' \cup \{y_1\}, \\ \mu'(T' \cup \{y_1\}, \mathbf{R}) &= T' \cup \{y_1, y_2\} \\ &\vdots \\ \mu'(T' \cup \{y_1, \ldots, y_{q-1}\}, \mathbf{R}) &= T' \cup \{y_1, \ldots, y_q\}. \end{aligned}$$

Note that $T' \cup \{y_1, \ldots, y_q\} = T_h - \{v_{i+1}\}$, and observe that the placements and removals defined above do not involve placement to the vertex v_{i+1}. Also, any vertex that is revisited in the new chain of moves would have been revisited anyway if the old chain of moves had been made. However, so far, the new chain does not revisit v_{i+1}, which was revisited in the old chain. To guarantee that v_{i+1} is not revisited in any future sequence of moves, we set $\mu'(S - \{v_{i+1}\}, \mathbf{R}) = \mu(S, \mathbf{R}) - \{v_{i+1}\}$ for any $S \subseteq V(H) \cup T_h$ and any $\mathbf{R}$ where $V(\mathbf{R}) \subseteq V(H)$. Otherwise, put $\mu'(S, \mathbf{R}) = \mu(S, \mathbf{R})$. We now have $c(\mu') < c(\mu)$, as required. □

We have shown the natural definitions of monotonicity to be equivalent, but is monotonicity important? Suppose we have r robbers in a tree T. We can modify program 1 so that, instead of letting T' be any component containing a robber, we set T' to be the component containing the ith robber, where i is minimal among those robbers who have not yet been caught; see program 2. This gives a program that catches the first robber (and any other robbers foolish enough to follow him), then

Search program 2. $\Pi(T, r)$ to capture r robbers in a tree T.

place(v), where v is any vertex of T.
Let $\mathbf{R} = [u_1 \dots u_r] \leftarrow$ *robbers_positions()*.
Let $T' \leftarrow T$.
While $V(\mathbf{R}) \cap V(T') \neq \emptyset$,
    Let i be minimal such that $u_i \neq *$.
    Let T' be the connected component of $T - v$ containing u_i,
        and let w be the (unique) vertex of T' adjacent to v.
    slide$(v \to w)$.
    Let $v \leftarrow w$.
    Let $\mathbf{R} \leftarrow$ *robbers_positions()*.
remove(v).

the second, and so on. There are two things to notice about this program: first, it is not monotone; second, it is winning against any number of robbers, without needing any more cops.

The same technique can be applied to transform any search program for one robber (on an arbitrary collection of graphs) into a nonmonotone program for any number of robbers. On the other hand, it is clear that monotonically searching for $r > 1$ robbers requires at least as many cops as does monotonically searching for a single robber. We summarize these observations in the following lemma.

LEMMA 5. *For any graph G and positive integer r,*

$$\begin{aligned} \mathbf{vams}(G, r) &= \mathbf{vams}(G, 1), \\ \mathbf{mvams}(G, r) &\geq \mathbf{mvams}(G, 1)\,. \end{aligned}$$

Thus, allowing nonmonotone strategies may make it easier to search for many robbers. This raises the question of what the cost of requiring monotonicity is when facing a crime wave. Given a graph G and $r \geq 1$, what is the ratio below?

$$\frac{\mathbf{mvams}(G, r)}{\mathbf{mvams}(G, 1)}.$$

In sections 5 and 6, we give a full answer for trees and an upper bound for general graphs. We postpone this until we have established some necessary results in the next section.

4. Invisible robbers. In this section we give brief descriptions of two game variants where the robbers are invisible and the cops must, therefore, determine their moves without reference to the robbers' position. In one of the variants we consider, the robber is active; in the other, he is lazy. Recall that an active robber can move at each round of the game, but a lazy robber may move only when a cop moves onto the vertex he occupies.

In both cases, as the robbers are now invisible, the game is no longer interactive and the cops' moves may be given in advance as a "predetermined" strategy. Thus, we define a *k-strategy* for k cops to be any consistent sequence $\mathcal{S} = S_0, S_1, \dots$ of sets in $V(G)^{[\leq k]}$.

Given such a strategy, we define the free space of an invisible, active robber to be the sequence

$$\begin{aligned} F_0 &= V(G), \\ F_{i+1} &= \{y \in V(G) - S_{i+1} \mid \text{there is an } (S_i, S_{i+1})\text{-avoiding} \\ &\qquad\qquad (x,y)\text{-path for some } x \in F_i\}. \end{aligned}$$

We say that $\mathcal{S}$ is *monotone* if $F_{i+1} \subseteq F_i$ for all $i \geq 0$ and $\mathcal{S}$ is *winning* if $F_i = \emptyset$ for some $i \geq 1$. Since the game is not interactive, we do not explicitly define plays, which will not feature in our analysis.

The *nonmonotone* and *monotone invisible active mixed search number* of a graph G are defined as follows:

$$\begin{aligned} \mathbf{iams}(G) &= \min\{k \mid \text{there exists a winning } k\text{-strategy on } G\}, \\ \mathbf{miams}(G) &= \min\{k \mid \text{there exists a monotone winning } k\text{-strategy on } G\}. \end{aligned}$$

It is known that $\mathbf{iams}(G) = \mathbf{miams}(G)$ [3]; that is, when searching for an invisible, active robber, insisting that the cops win the game monotonically does not increase the number of cops required. We also observe that searching for an active, invisible robber in a n-vertex graph G is equivalent to searching for n visible, active robbers. Intuitively, an invisible robber could be anywhere within his free space, while, with n robbers, there are plays in which every vertex of the free space really does contain a robber.

LEMMA 6. *For any graph G of order n,* $\mathbf{miams}(G) = \mathbf{mvams}(G, n)$.

Proof. We prove first that $\mathbf{miams}(G) \leq \mathbf{mvams}(G, n)$. Suppose we have a monotone winning (k, n)-strategy μ for the cops. Consider a μ-play $\mathcal{P} = S_0, \mathbf{R}_0, S_1, \mathbf{R}_1, \ldots$ where, for all $i \geq 0$, $V(\mathbf{R}_i) = F_i$. Such a play exists, since $V(R_0) = F_0 = V(G)$ and, by definition of F_{i+1}, it is possible for the robbers occupying the vertices F_i to move to the vertices in F_{i+1}. Notice, now, that the sequence $S_0, S_1, \ldots$ is a winning monotone k-strategy against an invisible, active robber.

For the converse, let $\mathcal{S} = S_0, S_1, \ldots$ be a monotone winning k-strategy against one invisible, active robber. We define a (k, n)-strategy μ by putting $\mu(S_i, \mathbf{R}) = S_{i+1}$ for any $i \geq 0$ and any $\mathbf{R} \in (V(G) \cup \{*\})^n$ with $V(\mathbf{R}) \subseteq F_{i+1}$. Any μ-play is monotone and winning because, no matter what moves the robbers make, $V(\mathbf{R}_i) \subseteq F_i$ for all i and the sequence $F_0, F_1, \ldots$ diminishes monotonically to the empty set. ∎

The case of an invisible, lazy robber is similar to the active case but with the difference that now, if the cops are at S and the robber at vertex v, then, when the cops move to S', the robber must stay at v unless $v \in S'$, in which case he can move along any (S, S')-avoiding path in the graph, as before. Thus, the robber moves only when a cop lands on his vertex.

We define free space, k-strategies, monotonicity, and winning against a lazy, invisible robber in the same way as in the active case and write $\mathbf{milms}(G)$ and $\mathbf{ilms}(G)$ for the corresponding *monotone* and *nonmonotone, invisible, lazy, mixed search number*, respectively.

LEMMA 7. *For any graph G,* $\mathbf{milms}(G) = \mathbf{mvams}(G, 1)$.

Proof. We prove first that $\mathbf{milms}(G) \leq \mathbf{mvams}(G, 1)$. Suppose we have a monotone winning $(k, 1)$-strategy μ for G, and let $\mathcal{E}$ be the set of the essential parts of all μ-plays. We can construct a monotone winning k-strategy against one invisible, lazy robber by taking an arbitrary concatenation of all of the sequences in $\mathcal{E}$.

We now show that $\mathbf{mvams}(G,1) \leq \mathbf{milms}(G)$. Let $\mathcal{S} = S_0, S_1, \ldots$ be a monotone winning k-strategy against one invisible, lazy robber. We will describe a search program against a visible, active robber. The first move is to place a cop on the vertex in S_1. Suppose that, at some stage, the cops occupy the vertices of some set S (not necessarily a set in $\mathcal{S}$). Let H be the connected component of $G - S$ that contains the robber, and let $S^* = \partial_G H$.

Now, let i be minimal such that S_i contains a vertex in H. We remove any cops that may be in $S - S_{i-1}$ and then play the move m that transforms S_{i-1} to S_i which, by definition of i, is not a removal. It is clear that we can play this move if it is a placement. If it is a slide, it must, by construction, be from a vertex in S^*. S must contain S^* by definition and, since $\mathcal{S}$ is a monotone strategy, S_{i-1} must also contain S^*: m is the first attack on H and the robber would be able to escape from that component if its boundary were not guarded.

This establishes that we have a monotone strategy. To see that it is winning, observe that, at each step, the robber's free space is decreased by at least one vertex (the target of m) so must, eventually, become empty. □

We define the *proper pathwidth* of a graph G to be $\mathbf{ppw}(G)$, the least k for which $G \preccurlyeq K_k \times P$ for some path P. (That is, $G \preccurlyeq G'$ for the graph G' formed from P by replacing the vertices with disjoint copies of K_k and adding a matching between the vertices of cliques corresponding to vertices adjacent in P.) Similarly, define the *proper treewidth* of a graph G as $\mathbf{ptw}(G)$, the least k for which $G \preccurlyeq K_k \times T$ for some tree T. It can be shown that $\mathbf{miams}(G) = \mathbf{ppw}(G)$ and $\mathbf{milms}(G) = \mathbf{ptw}(G)$ (see, e.g., [9, 12]).

COROLLARY 8. *For any graph G and any positive integer r,*

$$\mathbf{ptw}(G) \leq \mathbf{mvams}(G,r) \leq \mathbf{ppw}(G).$$

Proof. The proof is immediate from Lemmas 6 and 7 and the observation that, for any graph G of order n and any $r > n$, $\mathbf{mvams}(G,r) = \mathbf{mvams}(G,n)$, since the robbers can never occupy more than n distinct vertices. □

We could also consider multiple invisible robbers. We will not define the relevant games formally, but the informal remarks that follow should convince the reader that it would not be worth the effort to do so.

In the case of invisible, active robbers, it is clear that $\mathbf{miams}(G,r) = \mathbf{miams}(G)$ for any $r \geq 1$; essentially, no graph parameter defined through mixed search can ever be bigger than proper pathwidth. For invisible, lazy robbers, we must consider the conditions under which the robbers may move. The simplest scenario is that each robber may move only when a cop lands on the vertex he occupies. Define the *nonmonotone* and *monotone, invisible, lazy mixed search number* of a graph G to be, respectively, $\mathbf{ilms}(G,r)$ and $\mathbf{milms}(G,r)$, the least k such that there is a winning (respectively, monotone winning) k-strategy against r invisible, lazy robbers. In this case, it is not hard to see that $\mathbf{ilms}(G,r) = \mathbf{ilms}(G,1)$ (because, as usual, we can iterate the strategy for one robber to catch r robbers) and $\mathbf{milms}(G,r) = \mathbf{milms}(G,1)$ (the strategy used to prove Lemma 7 works as well for $r \geq 1$ lazy robbers as for one).

On the other hand, suppose that, when a cop lands on a vertex occupied by any robber, this fact is communicated to all the robbers, who may all move. Define the *nonmonotone* and *monotone, invisible, communicating, lazy mixed search numbers* of a graph G to be, respectively, $\mathbf{iclms}(G,r)$ and $\mathbf{miclms}(G,r)$, the least k such that there is a winning (respectively, monotone winning) k-strategy against r invisible,

Search program 3. $\Pi(T, v, r)$ to capture r robbers in a tree T monotonically.

place(v)
Let $\mathbf{R} \leftarrow$ *robbers_positions()*.
While $V(\mathbf{R}) \neq \emptyset$,
    Let $T_1, \ldots, T_\ell$ be the connected components of $T - v$
        containing at least one and at most $\lfloor \frac{r}{2} \rfloor$ robbers.
    For $i \in \{1, \ldots, \ell\}$,
        Choose any vertex $v_i \in V(T_i)$.
        Let r_i be the number of robbers in T_i.
        Call $\Pi(T_i, v_i, r_i)$.
    Let $\mathbf{R} \leftarrow$ *robbers_positions()*.
    if $V(\mathbf{R}) \cap V(T) \neq \emptyset$ (i.e., robbers remain in T), then
        Let T' be the unique connected component of $T - v$ where
            $V(\mathbf{R}) \subseteq V(T')$, and let w be the vertex of T'
            adjacent to v in T.
        slide$(v \to w)$.
        Let $v \leftarrow w$ and let $T \leftarrow T'$.
remove(v).

communicating, lazy robbers. We still have $\mathbf{iclms}(G, r) = \mathbf{ilms}(G)$ and, with just one robber, of course, $\mathbf{miclms}(G, 1) = \mathbf{milms}(G) = \mathbf{ptw}(G)$ since a single robber has nobody to communicate with. However, for any $r \geq 2$, having r invisible, communicating, lazy robbers is as bad as having an active, invisible robber: essentially, whenever the cops move to a vertex in the robbers' free space, they may disturb a robber, and, if they do, all the robbers may move. Thus, after any move, the cops must ensure that the entire boundary of the robbers' free space is guarded, just as in the active, invisible case. Hence, for $r \geq 2$, $\mathbf{miclms}(G, r) = \mathbf{miams}(G) = \mathbf{ppw}(G)$. We summarize these observations in the following theorem.

THEOREM 9. *For any graph G and any integers $r \geq 1$ and $s \geq 2$,*

$$\begin{aligned} \mathbf{ilms}(G, r) = \mathbf{iclms}(G, r) &= \mathbf{ilms}(G), \\ \mathbf{milms}(G, r) = \mathbf{miclms}(G, 1) &= \mathbf{milms}(G) = \mathbf{ptw}(G), \\ \mathbf{miclms}(G, s) = \mathbf{miams}(G, r) &= \mathbf{miams}(G) = \mathbf{ppw}(G)\,. \end{aligned}$$

We also note that it is believed but not yet proven[1] that $\mathbf{ilms}(G) = \mathbf{milms}(G)$.

5. Upper bounds. In this section, we demonstrate upper bounds for the value of $\mathbf{mvams}(G, r)$ for trees, in particular, and for all graphs.

LEMMA 10. *If T is a tree, then* $\mathbf{mvams}(T, r) \leq \lfloor \log r \rfloor + 1$.

Proof. Let $\Pi(T, r)$ be the search program that calls program 3 with v assigned to be any vertex of T.

We must prove that $\Pi(T, r)$ is winning and monotone and uses at most $\lfloor \log r \rfloor + 1$ cops. For this we use induction on the logarithm of the number of robbers. For the base case, notice that $\Pi(T, r)$ degenerates to program 1 when $r = 1$, as the program operates exclusively in the single component of $T - v$ containing $\lceil \frac{r}{2} \rceil = 1$ robber.

[1] A flawed proof appeared in [12].

Suppose that $\Pi(T, \lfloor \frac{r}{2} \rfloor)$ defines a winning, monotone $(q, \lfloor \frac{r}{2} \rfloor)$-strategy, where $q = \lfloor \log \frac{r}{2} \rfloor + 1 = \lfloor \log r \rfloor$. We now show that $\Pi(T, r)$ defines a winning, monotone $(q+1, r)$-strategy.

Before each slide move to w, each component of $T - v$ except for the one containing w contained at most $\lfloor \frac{r}{2} \rfloor$ robbers and has, by the inductive hypothesis, already been searched monotonically. Therefore, after each slide move, the free positions of the robbers have been updated from $V(T)$ to $V(T')$, where $V(T') \subset V(T)$. As the free positions for the robbers diminish, the program is monotone; and, as they diminish properly, the program is winning.

By the inductive hypothesis, each call to $\Pi(T_i, v_i, r_i)$ requires q cops. Meanwhile, there is only one additional cop in T (the cop on v), so $\Pi(T, r)$ uses $q + 1$ cops, as required. ∎

The following upper bound on $\mathbf{mvams}(T, r)$ is immediate from the previous lemma and Corollary 8.

COROLLARY 11. *For any tree T and for any positive integer r,*

$$\mathbf{mvams}(T, r) \leq \min \{\mathbf{ppw}(G), \lfloor \log r \rfloor + 1\}.$$

Our bound for trees leads to a bound for general graphs, obtained by considering tree decompositions.

THEOREM 12. *For any graph G and any positive integer r,*

$$\mathbf{mvams}(G, r) \leq \min \{\mathbf{ppw}(G), \mathbf{ptw}(G) \cdot (\lfloor \log r \rfloor + 1)\}.$$

Proof. Let $q = \lfloor \log r \rfloor + 1$.

By Corollary 8, it is enough to show that $\mathbf{mvams}(G, r) \leq \mathbf{ptw}(G) \cdot q$. Assuming that $\mathbf{ptw}(G) \leq k$, we have $G \preccurlyeq G' = K_k \times T$ for some tree T. We assume that the vertices of the clique in G' corresponding to the vertex $v \in T$ are $K(v) = \{v_1, \ldots, v_k\}$ and, for every edge $uv \in T$, the corresponding edges in G' are $u_1v_1, \ldots, u_kv_k$. For each $S \subseteq V(T)$, let $K(S) = \bigcup_{v \in S} K(v)$.

By Lemma 10, $\mathbf{mvams}(T, r) \leq q$, so there is a monotone winning (q, r)-strategy μ for T. We use μ to construct a monotone, winning (kq, r)-strategy μ' for G'. The idea is that we simulate a single cop on $v \in T$ with k cops, one on each vertex of $K(v) \subseteq G'$. Each placement, removal and slide is replaced by the equivalent operation on each of these k cops in turn.

Formally, let $S \in V(T)^{[\leq q]}$, let $\mathbf{R} \in (V(T) \cup \{*\})^r$, and let $S' = \mu(S, \mathbf{R})$. There are three cases, depending on the type of the move from S to S'.

Placement. Let $\{v\} = S' - S$. For any $j \in \{1, \ldots, k\}$, let $S_j = K(S) \cup \{v_1, \ldots, v_{j-1}\}$. For any $\mathbf{R}'$ with $V(\mathbf{R}') \subseteq K(V(\mathbf{R})) - \{v_1, \ldots, v_{j-1}\}$, we set $\mu'(S_j, \mathbf{R}') = S_j \cup \{v_j\}$.

Removal. Let $\{v\} = S - S'$. For any $j \in \{1, \ldots, k\}$, let $S_j = K(S) - \{v_1, \ldots, v_{j-1}\}$. For any $\mathbf{R}'$ with $V(\mathbf{R}') \subseteq K(V(\mathbf{R})) \cup \{v_1, \ldots, v_{j-1}\}$, we set $\mu'(S_j, \mathbf{R}') = S_j - \{v_j\}$.

Sliding. Let $\{v\} = S - S'$ and $\{w\} = S' - S$. For any $j \in \{1, \ldots, k\}$, let $S_j = \big(K(S) - \{v_1, \ldots, v_{j-1}\}\big) \cup \{w_1, \ldots, w_{j-1}\}$. For any $\mathbf{R}'$ with $V(\mathbf{R}') \subseteq \big(K(V(\mathbf{R})) \cup \{v_1, \ldots, v_{j-1}\}\big) - \{w_1, \ldots, w_{j-1}\}$, we set $\mu'(S_j, \mathbf{R}') = (S_j - \{v_j\}) \cup \{w_j\}$.

Notice that the fact that μ is winning and monotone implies the same for μ'. Moreover, in each μ'-play any set $S \in V(T)^{[\leq q]}$ corresponds to a sequence of k sets in $V(G')^{[\leq kq]}$. Therefore, $\mathbf{mvams}(G', r) \leq kq$ and thus $\mathbf{mvams}(G, r) \leq kq$, by Proposition 3. ∎

6. Lower bounds. We now give lower bounds for $\mathbf{mvams}(T, r)$ for trees T. We introduce a general form of graph composition and analyze the search numbers of graphs formed by such compositions. We define the composition for general graphs, though our main use of the construction will be for trees.

We say that graphs $G_0, \ldots, G_3$ are *k-connectable* if

- for $0 \leq i \leq 3$, G_i is k-connected;
- G_1, G_2, and G_3 are pairwise disjoint;
- for $1 \leq i \leq 3$, $|U_i| = k$, where $U_i = V(G_0) \cap V(G_i)$; and
- for $1 \leq i < j \leq 3$ and $h \in \{1, 2, 3\} - \{i, j\}$, $G_0 - U_h$ contains a set P_{ij} of k pairwise vertex-disjoint paths from U_i to U_j.

Note that the k-connectedness of G_0 already implies the existence in that graph of a set of k pairwise vertex-disjoint paths from U_i to U_j, but these do not necessarily avoid U_h.

Let G be a graph, with $x \in V(G)$ and $U \subseteq V(G) - \{x\}$. An *(x, U)-fan* is a set of paths in G, one from x to each vertex in U, where the paths are pairwise disjoint, except for the common endpoint x.

LEMMA 13. *Let $G_0, \ldots, G_3$ be k-connectable. Then $G = G_0 \cup \cdots \cup G_3$ is k-connected.*

Proof. It suffices to show that, for any $u, v \in V(G)$, G contains k independent (u, v)-paths. If $u, v \in V(G_i)$ for some i, then the result follows immediately from the k-connectedness of G_i. So, suppose that there are $i < j$ such that $u \in V(G_i) - V(G_j)$ and $v \in V(G_j) - V(G_j)$. There are two cases.

If $i = 0$, then, by [4, Theorem 2.6], G_0 contains a (u, U_j) fan and G_j contains a (v, U_j) fan. Since $V(G_0) \cap V(G_j) = U_j$, these fans are disjoint, except for the vertices in U_j. Their union is, therefore, a set of k independent (u, v)-paths in G.

If $i > 0$ then, as above, G_i contains a (u, U_i)-fan and G_j contains a (u, U_j)-fan and these fans are disjoint. The union of the two fans and the paths P_{ij} is a set of k independent (u, v)-paths in G. ∎

The following is the key technical result of this section.

LEMMA 14. *Let $G_0, \ldots, G_3$ be k-connectable, with $\mathbf{mvams}(G_i, \lfloor \frac{r}{2} \rfloor) \geq q$ for each i, and let $G = G_0 \cup \cdots \cup G_3$. Then $\mathbf{mvams}(G, r) \geq q + k$.*

Proof. Suppose, toward a contradiction, that $\mathbf{mvams}(G, r) < q + k$, and let μ be a smooth, monotone, winning $(q + k - 1, r)$-strategy for G. By Lemma 13, G is k-connected. Hence, whenever there are robbers in the graph and the position of the cops is S, with $|S| < k$, the robbers' free space is the whole of $G - S$. Let $\mathcal{P} = S_0, \mathbf{R}_0, S_1, \mathbf{R}_1, \ldots$ be a play, and let α be minimal such that

- the free space of some robber in $\mathbf{R}_\alpha$ does not include the whole of $V(G_i) - S_\alpha$ for some $i \geq 1$, or
- $|V(G_i) \cap S_\alpha| \geq k$ for some $i \geq 1$.

α is well defined because μ is winning. We may assume that the robbers, who are aware of the cops' strategy μ, arrange to maximize α. By smoothness, the cops make the same moves in all plays up to (and including) the αth move, as long as the robbers behave as we have described.

As argued above, $|S_\alpha| \geq k$; otherwise, the free space of every robber is just $V(G) - S_\alpha$ and no G_i contains k cops. We may assume, without loss of generality, that $|V(G_1) \cap S_\alpha| < k$ and $|V(G_2) \cap S_\alpha| < k$. (If necessary, rename the parts to achieve this.) Let $S'_\alpha = S_\alpha - (V(G_1) \cap V(G_2))$. By construction, $|S'_\alpha| \geq k$.

At the point when the cops make move α, the free space of every robber includes all of $V(G_1) - S_{\alpha-1}$ and all of $V(G_2) - S_{\alpha-1}$. Therefore, we may assume that $\mathbf{R}_\alpha$ has

$\lceil \frac{r}{2} \rceil$ robbers in G_1 and $\lfloor \frac{r}{2} \rfloor$ robbers in G_2. We will show that the assumption that μ is monotone and winning contradicts the hypothesis that $\mathbf{mvams}(G_1, \lfloor \frac{r}{2} \rfloor) \geq q$ and $\mathbf{mvams}(G_2, \lfloor \frac{r}{2} \rfloor) \geq q$.

Let T_μ be the labeled tree representing the strategy μ. We may delete from T_μ any subtree whose root has an incoming edge labeled $\mathbf{R}$ for some $\mathbf{R}$ containing more than $\lceil \frac{r}{2} \rceil$ robbers in G_1 or more than $\lfloor \frac{r}{2} \rfloor$ robbers in G_2. This restriction on the robbers' position can only make the game easier for the cops.

First, let T_1 be the tree formed from T_μ by deleting every vertex outside $V(G_1)$ from every vertex label and replacing every vertex outside $V(G_1)$ with $*$ in every edge label. These deletions from the labels may result in two adjacent vertices u and v having identical labels (because the corresponding move in T_μ was outside G_1) and two edges from the same vertex having identical labels (because a move within G_1 in T_μ depended on the positions of robbers outside G_1). As such, T_1 is nondeterministic, in the sense discussed in section 2.

Call the subtree of T_1 rooted at vertex x *bad* if either the label of x is a set of size at least q or there is some $\mathbf{R}$ such that, whenever the edge (x, y) is labeled $\mathbf{R}$, the subtree rooted at y is bad. Call a subtree *good* if it is not bad. We claim that T_1 is, itself, bad. Suppose not. Delete all bad subtrees from T_1 to give T_1'. Because every vertex of the resulting tree has at least one child for each possible $\mathbf{R}$, T_1' defines a winning $(q-1, \lceil \frac{r}{2} \rceil)$-strategy for G_1, contradicting the hypothesis that $\mathbf{mvams}(G_1, \lceil \frac{r}{2} \rceil) \geq \mathbf{mvams}(G_1, \lfloor \frac{r}{2} \rfloor) \geq q$. (In fact, the strategy defined by T_1' may be nondeterministic, but we may take an arbitrary deterministic restriction.)

Now, let T_1^* be the tree that results from deleting all good subtrees from T_1. T_1^* is a tree of all plays where the robbers force there to be q cops in G_1 and can be seen as a certificate of the fact that no monotone winning $(q+k-1, r)$-strategy for G can induce a monotone winning $(q-1, \lceil \frac{r}{2} \rceil)$-strategy for G_1.

Let T be the subtree of T_μ consisting of those vertices in T_1^*, and let T_2 be the tree made from T by deleting every vertex outside $V(G_2)$ from every vertex label and replacing every vertex outside $V(G_2)$ with $*$ in every edge label. T_2 defines a nondeterministic strategy for G_2 in the same sense that T_1 does for G_1. Because T_1 is a strategy for G_1, it includes responses for every possible position $\mathbf{R}$ of the robbers within that subgraph. In turn, every position of robbers in G whose restriction to G_1 is $\mathbf{R}$ will produce the same response within T_1': in particular, then, T_2 contains a vertex corresponding to this position.

Using the hypothesis that $\mathbf{mvams}(G_2, \lfloor \frac{r}{2} \rfloor) \geq q$ and the same argument as for G_1, we see that T_2 is also bad. Again, define T_2^* by deleting all good subtrees from T_2. By construction, any path in T_2^* corresponds to a path in T_1^* and a path in T_μ. Choose any such path, and let $\mathcal{P} = S_0, \mathbf{R}_0, S_1, \mathbf{R}_2, \ldots$ be the corresponding μ-play. For $i \in \{1, 2\}$, let $\mathcal{P}^i = S_0^i, \mathbf{R}_0^i, S_1^i, \mathbf{R}_2^i, \ldots$ be the labels of the corresponding path in T_i^*. Note that $S_j^1 \cup S_j^2 \subseteq S_j$ and $V(\mathbf{R}_j^1) \cup V(\mathbf{R}_j^2) \subseteq V(\mathbf{R}_j)$ for all $j \geq 0$.

For $i \in \{1, 2\}$, let c_i be minimal such that $V(\mathbf{R}_{c_i}^i) = \emptyset$. Since no move of the cops can simultaneously capture robbers in both G_1 and G_2, we must have $c_1 \neq c_2$. Without loss of generality, we may assume $c_1 < c_2$. Let h be minimal such that $|S_h^1| \geq q$. Since μ is a $(q+k-1, r)$-strategy, we must have $|S_h^2| < k$, and, further, at least one of the cops originally placed on $v \in S_\alpha'$ must have been removed. This contradicts the monotonicity of μ because, after move α of the game, no robber in G_2 could reach vertex v, but now they all can. ∎

A 3-*star composition* of disjoint, connected graphs G_1, G_2, and G_3 is any graph $\curlyvee(G_1, G_2, G_3)$ formed by adding a new vertex v to $G_1 \cup G_2 \cup G_3$ and adding one edge

from v to each of the three component graphs. Observe that the 3-star composition of G_1, G_2, and G_3 is a special case of the graphs $K_{1,3}$, G_1, G_2, and G_3 being 1-connectable. Hence, the following is an immediate corollary of Lemma 14.

COROLLARY 15. *Let $G = \mathsf{Y}(G_1, G_2, G_3)$, where, for each $i \in \{1,2,3\}$, it holds that $\mathbf{mvams}(G_i, \lfloor \frac{r}{2} \rfloor) \geq q$. Then, $\mathbf{mvams}(G, r) > q$.*

We are now ready to show that the upper bound of Corollary 11 is, in fact, an exact characterization of $\mathbf{mvams}(T, r)$ for all trees T and natural numbers r.

THEOREM 16. *For any tree T and $r \geq 1$,*

$$\mathbf{mvams}(T, r) = \min\{\mathbf{ppw}(T), \lfloor \log r \rfloor + 1\}.$$

Proof. By Corollary 11, it suffices to show that $\mathbf{mvams}(T, r) \geq \min\{\mathbf{ppw}(T), \lfloor \log r \rfloor + 1\}$. For this, we use induction on $q = \lfloor \log r \rfloor + 1$. For the base case, $q = 1$, program 1 shows that $\mathbf{mvams}(T, r) = 1 = q$. Suppose the result holds for all values smaller than q, and let T be a tree. If $\mathbf{ppw}(T) = 1$, then T is a path and $\mathbf{mvams}(T, r) = 1$ for any r, as required. Otherwise, it is known from [14] that we can write $T = \mathsf{Y}(T_1, T_2, T_3)$, where, for each i, $\mathbf{ppw}(T_i) = \mathbf{ppw}(T) - 1$. By the inductive hypothesis, $\mathbf{mvams}(T_i, \lfloor \frac{r}{2} \rfloor) \geq \min\{\mathbf{ppw}(T) - 1, q - 1\}$. By Corollary 15, $\mathbf{mvams}(T, r) \geq \min\{\mathbf{ppw}(T) - 1, q - 1\} + 1 = \min\{\mathbf{ppw}(T), q\}$, as required. ∎

We do not have a lower bound for $\mathbf{mvams}(G, r)$ for general graphs. However, we are able to demonstrate that the upper bound of Theorem 12 is reached by an infinite class of graphs. Toward this end, define the parameterized graph class $\mathcal{O}_w$ recursively as follows: $\mathcal{O}_0 = \{K_1\}$, and $G \in \mathcal{O}_{w+1}$ if and only if G is a 3-star composition of three graphs in $\mathcal{O}_w$. From [13], $\mathcal{O}_w$ contains all minor-minimal trees with proper pathwidth at least w. Define

$$\mathcal{O}_w^k = \{T \times K_k : T \in \mathcal{O}_w\}.$$

It follows from Lemma 14 that the upper bound of Theorem 12 is tight as the bound is attained by all graphs in $\mathcal{O}_w^k$. Further, because the graphs in $\mathcal{O}_w^k$ are minor-minimal, the bound of Theorem 12 is attained by all products of trees and cliques.

We have remarked that k-composition is a generalization of 3-star composition. Finally, we show that the above results on proper pathwidth of 3-star compositions of graphs can be strengthened to k-compositions.

COROLLARY 17. *Let $G_0, \ldots, G_3$ be k-connectable graphs, each of proper pathwidth at least w, and let $G = G_0 \cup \cdots \cup G_3$. Then, $\mathbf{ppw}(G) \geq w + k$.*

Proof. Let $n = |V(G)|$.

$$\begin{aligned}
\mathbf{ppw}(G) &= \mathbf{mvams}(G, 2n) && \text{(Corollary 8)} \\
&\geq \min_{0 \leq i \leq 3} \mathbf{mvams}(G_i, n) + k && \text{(Lemma 14)} \\
&= \min_{0 \leq i \leq 3} \mathbf{ppw}(G_i) + k && \text{(Corollary 8)} \\
&\geq w + k. && \blacksquare
\end{aligned}$$

7. Conclusions and open problems. We have presented our results in the setting of mixed search (i.e., searching with placement, removal, and sliding of cops). For node search (searching with only placement and removal of cops), we can similarly define parameters $\mathbf{vans}(G)$ and $\mathbf{mvans}(G, r)$ for the general and monotone node search numbers for r visible, active robbers. Similarly, we can adapt all definitions of mixed-search parameters given in this paper to their node search counterparts.

The difference between mixed search and node search is not very great: node search can be reduced to mixed search, and the node search number is either equal to the corresponding mixed search number or one greater, depending on the graph.

We could, in principle, rewrite the present paper in terms of node search. Writing $\mathbf{pw}(G)$ and $\mathbf{tw}(G)$ for the well-known parameters of pathwidth and treewidth, it can be shown, using the results in [6, 11], that, for a graph of order n, $\mathbf{vians}(G,1) = \mathbf{tw}(G)+1$ and $\mathbf{vians}(G,n) = \mathbf{pw}(G)+1$. For completeness, we restate our core results for this setting:

$$\begin{aligned}
&\mathbf{mvans}(T,r) = \min\{\mathbf{pw}(T), \lfloor \log r \rfloor + 1\} + 1 && \text{(for any tree } T\text{)},\\
&\mathbf{mvans}(G,r) \le \min\{\mathbf{pw}(G)+1, (\mathbf{tw}(G)+1)\cdot(\lfloor \log r \rfloor + 1)\} && \text{(for any graph } G\text{)}.
\end{aligned}$$

Moreover, the framework of this paper can be applied to the other classical search variant, edge search. As this version can also be reduced to mixed searching (see, e.g., [3, 12]), we make no further comments in this direction.

The problem settled in this paper can be stated in the following way: given a graph G, what is the maximum number of visible, active robbers that can be captured by k cops? According to our results, this number is unbounded if $k \ge \mathbf{ppw}(G)$. In the case that $k < \mathbf{ppw}(G)$, the maximum number of robbers that can be caught in a tree is 2^{k-1}, and, for general graphs, it is at least $2^{k/\mathbf{ptw}(G)-1}$. This interpretation of our results may be useful for estimating how many sweeps of a graph a small number of cops needs to catch a large number of robbers.

We identify three main open problems on the study of graph searching for many robbers. The first is to find good lower bounds for $\mathbf{mvams}(G,r)$ in terms of G and r, for general graphs, corresponding to the bounds for trees found in this paper. We believe that this is a hard task as such a study appears to require the identification of obstructions for $\mathbf{mvams}(G,r)$ for all values of r.

Another open problem is to find graph decompositions corresponding to the game, tuning between (proper) tree decompositions (the case for one robber) and (proper) path decompositions (one robber per vertex). It is unclear what form such a family of decompositions would take.

Finally, it would be interesting to know whether there is any relation between our results and the search game defined by Fomin, Fraigniaud, and Nisse [8]. That game has only one robber but tunes between pathwidth and treewidth by limiting the number of rounds at which the cops may ask for the position of the robber. This provides an alternative way of tuning the interactivity of the game: it is fully interactive if the cops may ask for the robber's position at every move and fully predetermined if they may never ask for his position. Correspondingly, our game is fully interactive with a single robber and fully predetermined with a robber for each vertex of the graph.

REFERENCES

[1] B. Alspach, *Searching and sweeping graphs: A brief survey*, Matematiche (Catania), 59 (2004), pp. 5–37 (2006).

[2] D. Bienstock, *Graph searching, path-width, tree-width, and related problems (a survey)*, in Reliability of Computer and Communication Networks (New Brunswick, NJ, 1989), DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 5, AMS, Providence, RI, 1991, pp. 33–49.

[3] D. Bienstock and P. Seymour, *Monotonicity in graph searching*, J. Algorithms, 12 (1991), pp. 239–245.

[4] B. Bollobás, *Extremal Graph Theory*, Dover, Mineola, NY, 2004. Reprint of the 1978 original.

[5] Y. Colin de Verdière, *Multiplicities of eigenvalues and tree-width of graphs*, J. Combin. Theory Ser. B, 74 (1998), pp. 121–146.
[6] N. D. Dendris, L. M. Kirousis, and D. M. Thilikos, *Fugitive-search games on graphs and related parameters*, Theoret. Comput. Sci., 172 (1997), pp. 233–254.
[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] F. V. Fomin, P. Fraigniaud, and N. Nisse, *Nondeterministic graph searching: From pathwidth to treewidth*, in Mathematical Foundations of Computer Science 2005, Lecture Notes in Comput. Sci. 3618, Springer-Verlag, Berlin, 2005, pp. 364–375.
[9] F. V. Fomin and D. M. Thilikos, *Multiple Edges Matter When Searching a Graph: The Exact Figure*, manuscript.
[10] P. Hunter and S. Kreutzer, *Digraph measures: Kelly decompositions, games, and orderings*, in Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (New Orleans, 2007), ACM, New York, SIAM, Philadelphia, 2007, pp. 637–644.
[11] P. D. Seymour and R. Thomas, *Graph searching and a min-max theorem for tree-width*, J. Combin. Theory Ser. B, 58 (1993), pp. 22–33.
[12] Y. C. Stamatiou and D. M. Thilikos, *Monotonicity and inert fugitive search games*, Electronic Notes in Discrete Mathematics, 3 (1999).
[13] A. Takahashi, S. Ueno, and Y. Kajitani, *Minimal acyclic forbidden minors for the family of graphs with bounded path-width*, Discrete Math., 127 (1994), pp. 293–304.
[14] A. Takahashi, S. Ueno, and Y. Kajitani, *Mixed searching and proper-path-width*, Theoret. Comput. Sci., 137 (1995), pp. 253–268.
[15] D. M. Thilikos, *Algorithms and obstructions for linear-width and related search parameters*, Discrete Appl. Math., 105 (2000), pp. 239–271.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 369–371

REDUCTION OF ROTA'S BASIS CONJECTURE TO A PROBLEM ON THREE BASES*

TIMOTHY Y. CHOW†

Abstract. It is shown that Rota's basis conjecture follows from a similar conjecture that involves just three bases instead of n bases.

Key words. common independent sets, non–base-orderable matroid, odd wheel

AMS subject classifications. Primary, 05B20; Secondary, 15A03

DOI. 10.1137/080723727

1. Introduction. In 1989, Rota formulated the following conjecture, which remains open.

CONJECTURE 1 (Rota's basis conjecture). *Let M be a matroid of rank n on n^2 elements that is a disjoint union of n bases $B_1, B_2, \ldots, B_n$. Then there exists an $n \times n$ grid G containing each element of M exactly once, such that for every i the elements of B_i appear in the ith row of G and such that every column of G is a basis of M.*

Partial results toward this conjecture may be found in [1, 2, 3, 4, 5, 6, 7, 8, 12, 14, 15]. Now consider the following conjecture.

CONJECTURE 2. *Let M be a matroid of rank n on $3n$ elements that is a disjoint union of 3 bases. Let $I_1, I_2, \ldots, I_n$ be disjoint independent sets of M, with $0 \le |I_i| \le 3$ for all i. Then there exists an $n \times 3$ grid G containing each element of M exactly once, such that for every i the elements of I_i appear in the ith row of G and such that every column of G is a basis of M.*

The main purpose of the present note is to make the following observation.

THEOREM 3. *Conjecture* 2 *implies Conjecture* 1.

Our proof is inspired by the proof of Theorem 4 in [10].

Proof. Since Conjecture 1 is known if $n \le 2$, we may assume that $n \ge 3$. Let M be given as in the hypothesis of Conjecture 1. Define a *transversal* to be a subset $\tau \subseteq M$ that contains exactly one element from each B_i. Define a *double partition* of M to be a pair (β, τ) where $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$ is a partition of M into n pairwise disjoint bases β_i and $\tau = (\tau_1, \tau_2, \ldots, \tau_n)$ is a partition of M into n pairwise disjoint transversals. Given a double partition (β, τ), define

$$\mu(\beta, \tau) = \sum_{i \ne j} |\beta_i \cap \tau_j|.$$

Observe that if $\mu(\beta, \tau) = 0$, then necessarily $\beta_i = \tau_i$ for all i, and then Rota's basis conjecture follows—just let the (i, j)th entry of G be $B_i \cap \tau_j$.

So let (β, τ) be an arbitrary double partition with $\mu(\beta, \tau) > 0$. We show how to construct a double partition (β', τ') with $\mu(\beta', \tau') < \mu(\beta, \tau)$; the proof is then complete, by infinite descent, since by hypothesis there exists at least one double partition. Since $\mu(\beta, \tau) > 0$, there exist β_i and τ_j with $i \ne j$ such that $\beta_i \cap \tau_j \ne \emptyset$. Since $n \ge 3$, there also exists k such that i, j, and k are all distinct. It will simplify

*Received by the editors May 8, 2008; accepted for publication (in revised form) September 29, 2008; published electronically January 7, 2009.
http://www.siam.org/journals/sidma/23-1/72372.html
†Center for Communications Research, 805 Bunn Drive, Princeton, NJ 08540 (tchow@mit.edu).

notation to assume that $i = 1$, $j = 2$, and $k = 3$; no generality is lost, and it will be convenient to be able to reuse the index variables i and j below. Let $S = \beta_1 \cup \beta_2 \cup \beta_3$, let $T = \tau_1 \cup \tau_2 \cup \tau_3$, and let $M' = M|S$ (i.e., M restricted to the ground set S).

For each i, let $I_i = B_i \cap T \cap S$. Then I_i is an independent subset of the matroid M', and $|I_i| \leq |B_i \cap T| \leq 3$. The I_i are pairwise disjoint because the B_i are pairwise disjoint. Therefore, we may apply Conjecture 2 to obtain an $n \times 3$ grid G' whose columns β'_1, β'_2, and β'_3 are disjoint bases of M' (and therefore are bases of M) and whose ith row contains the elements of I_i.

To construct the desired double partition (β', τ'), let $\beta' = \beta$ except with β_1, β_2, and β_3 replaced with β'_1, β'_2, and β'_3, respectively. Similarly, let $\tau' = \tau$ except with τ_1, τ_2, and τ_3 replaced with τ'_1, τ'_2, and τ'_3, which are defined as follows. Let G'' be any $n \times 3$ grid whose ith row contains the elements of $B_i \cap T$ in some order, and whose (i, j)th entry agrees with that of G' whenever that entry is in I_i. Clearly G'' exists (though it may not be unique). Let τ'_j be the jth column of G'' for $j = 1, 2, 3$.

It is easily verified that what we have done is to regroup the elements of M' into three new bases and to regroup the elements of T into three new transversals in such a way that the contribution to $\mu(\beta', \tau')$ from intersections of the new bases with the new transversals is reduced to zero, and such that the total of the other contributions to μ is unchanged. Thus the overall value of μ is reduced, as required. □

Careful inspection of the above proof shows that it is easily adapted to prove a stronger statement than Theorem 3. Let $C(k)$ denote the statement obtained by replacing "3" with "k" throughout Conjecture 2. Then the above argument, mutatis mutandis, yields the following result.

THEOREM 4. *For any $\ell \geq k \geq 2$, $C(k)$ implies $C(\ell)$.*

In particular, proving $C(k)$ for any fixed k would prove Rota's basis conjecture (in fact, a stronger statement, namely, $C(n)$) for all n greater than or equal to that fixed k.

It is therefore natural to ask why we have formulated Conjecture 2 as $C(3)$ rather than as $C(2)$. The reason is that $C(2)$ is false. The simplest counterexample is a well-known stumbling block that is partly responsible for the fact that there is no known general "matroid union intersection theorem," i.e., a criterion for determining the minimum number of common independent sets that a set with two matroid structures on it can be partitioned into. Namely, take $M(K_4)$, the graphic matroid of the complete graph on four vertices, and let the I_i be the three pairs of nonincident edges of K_4. Another counterexample arises from a matroid that Oxley [11] calls J. Representing J by vectors in Euclidean 4-space, we can, for example, let

$$
\begin{aligned}
I_1 &= \{(-2, 3, 0, 1), (0, 0, 1, 1)\},\\
I_2 &= \{(0, 2, 0, 1), (1, 0, 3, 1)\},\\
I_3 &= \{(1, 0, 0, 1), (0, 1, 2, 1)\},\\
I_4 &= \{(0, 1, 0, 1), (4, 0, 0, 1)\}.
\end{aligned}
$$

It may be possible to construct other examples from non–base-orderable matroids such as those in [9].

Despite these counterexamples to $C(2)$, we believe that Conjecture 2 is plausible. Using a database of matroids with nine elements kindly supplied by Gordon Royle [13], we have computationally verified Conjecture 2 for the case $n = 3$.

In an earlier version of this paper, the formulation of Conjecture 2 did not require the I_i to be independent. A counterexample to that version of the conjecture was

found by Colin McDiarmid. Take the complete graph on the vertex set $\{1, 2, 3, 4\}$, and create an extra copy of the three edges incident to vertex 4. Call the edges $12, 13, 14, 23, 24, 34, 14', 24', 34'$, and let $I_1 = \{14, 14', 23\}$, $I_2 = \{24, 24', 13\}$, and $I_3 = \{34, 34', 12\}$. More generally, as pointed out by an anonymous referee, if k is odd, then a wheel with $k - 1$ copies of each of its k spokes yields a counterexample to $C(k)$ if the I_i are not required to be independent.

In closing, we speculate that Conjecture 2 might be provable using the following strategy. First, develop a modified version of $C(2)$ that says that the conclusion holds provided certain "obstructions" (such as $M(K_4)$ and J) are absent. Then use Rado's theorem (12.2.2 of [11]), or a suitable strengthening of it, to construct a first column of G in such a way that the remaining $2n$ elements are obstruction-free. Applying the modified version of $C(2)$ would then yield the desired result. The analysis of obstructions should hopefully be tractable since there are only three columns to consider.

Acknowledgments. I wish to thank Jonathan Farley, Patrick Brosnan, and James Oxley for useful discussions, and a referee for correcting an error in my counterexample based on J.

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), R28.
[6] J. Geelen and P. J. Humphries, *Rota's basis conjecture for paving matroids*, SIAM J. Discrete Math., 20 (2006), pp. 1042–1045.
[7] J. Geelen and K. Webb, *On Rota's basis conjecture*, SIAM J. Discrete Math., 21 (2007), pp. 802–804.
[8] 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.
[9] A. W. Ingleton, *Non-base-orderable matroids*, in Proceedings of the 5th British Combinatorial Conference, Aberdeen, Scotland, 1975, pp. 355–359.
[10] J. Keijsper,*An algorithm for packing connectors*, J. Combin. Theory Ser. B, 74 (1998), pp. 397–404.
[11] J. G. Oxley, *Matroid Theory,* Oxford University Press, Oxford, UK, 1992.
[12] V. Ponomarenko, *Reduction of jump systems*, Houston J. Math., 30 (2004), pp. 27–33.
[13] D. Mayhew and G. F. Royle, *Matroids with nine elements*, J. Combin. Theory Ser. B, 98 (2008), pp. 415–431.
[14] M. Wild, *On Rota's problem about n bases in a rank n matroid*, Adv. Math., 108 (1994), pp. 336–345.
[15] P. Zappa, *The Cayley determinant of the determinant tensor and the Alon–Tarsi conjecture*, Adv. in Appl. Math., 19 (1997), pp. 31–44.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 372–383

6-CRITICAL GRAPHS ON THE KLEIN BOTTLE*

KEN-ICHI KAWARABAYASHI†, DANIEL KRÁL'‡, JAN KYNČL§, AND BERNARD LIDICKÝ¶

Abstract. We provide a complete list of 6-critical graphs that can be embedded on the Klein bottle settling a problem of Thomassen [*J. Combin. Theory Ser. B*, 70 (1997), pp. 67–100, Problem 3]. The list consists of nine nonisomorphic graphs which have altogether 18 nonisomorphic 2-cell embeddings and one embedding that is not 2-cell.

Key words. graphs on surfaces, 6-critical graphs, Klein bottle, Heawood formula

AMS subject classifications. 05C15, 05C10

DOI. 10.1137/070706835

1. Introduction. We study colorings of graphs embedded on surfaces. It is well known [13] that the chromatic number of a graph embedded on a surface of Euler genus g is bounded by the Heawood number $H(g) = \lfloor \frac{7+\sqrt{24g+1}}{2} \rfloor$. The Dirac map color theorem [5, 6] asserts that a graph G embedded on a surface of Euler genus $g \neq 0, 2$ is $(H(g)-1)$-colorable unless G contains a complete graph of order $H(g)$ as a subgraph. Dirac's theorem can be rephrased using the language of critical graphs as follows: the only $H(g)$-critical graph that can be embedded on a surface of Euler genus $g \neq 0, 2$ is the complete graph of order $H(g)$. Recall that a graph G is k-critical if G is k-chromatic and every proper subgraph of G is $(k-1)$-colorable.

In fact, Dirac [5] showed that there are only finitely many k-critical graphs, $k \geq 8$, that can be embedded on a fixed surface. The number of 7-critical graphs that can be embedded on a fixed surface is also finite by classical results of Gallai [11, 12] as pointed out by Thomassen in [16]. Later, Thomassen [18] established that the number of 6-critical graphs that can be embedded on any fixed (orientable or nonorientable) surface is finite (see also [10] for related results on 7-critical graphs). This result is best possible as there are infinitely many k-critical graphs, $3 \leq k \leq 5$, that can be embedded on any fixed surface different from the plane [9].

In this paper, we focus on 6-critical graphs on surfaces, motivated by Problem 3 from [18]. As every plane graph is 4-colorable [1, 2, 14], there are no 6-critical graphs

*Received by the editors October 30, 2007; accepted for publication (in revised form) September 9, 2008; published electronically January 14, 2009. A significant part of these results was obtained during a DIMACS-DIMATIA REU project of Jan Kynčl and Bernard Lidický (supervised by Daniel Král') during their stay at DIMACS in June 2007. Their work was partially supported by KONTAKT grant ME 886.

http://www.siam.org/journals/sidma/23-1/70683.html

†National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, Japan (k_keniti@nii.ac.jp). This author's research was partly supported by Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research, by C & C Foundation, by Kayamori Foundation, and by Inoue Research Award for Young Scientists.

‡Institute for Theoretical Computer Science (ITI), Faculty of Mathematics and Physics, Charles University, Malostranské náměstí 25, 118 00 Prague 1, Czech Republic (kral@kam.mff.cuni.cz). ITI is supported as project 1M0545 by the Czech Ministry of Education.

§Department of Applied Mathematics and Institute for Theoretical Computer Science (ITI), Faculty of Mathematics and Physics, Charles University, Malostranské náměstí 25, 118 00 Prague 1, Czech Republic (kyncl@kam.mff.cuni.cz).

¶Department of Applied Mathematics, Faculty of Mathematics and Physics, Charles University, Malostranské náměstí 25, 118 00 Prague 1, Czech Republic (bernard@kam.mff.cuni.cz).

in the plane. The Dirac map color theorem implies that the complete graph of order six is the only 6-critical projective planar graph. Thomassen [16] classified 6-critical toroidal graphs: the only 6-critical graphs that can be embedded on the torus are the complete graph K_6, the join of the cycles C_3 and C_5 (recall that the join of two graphs G_1 and G_2 is the graph obtained by adding all edges between G_1 and G_2), the graph obtained by applying Hajos's construction to two copies of K_4 and then by adding K_2 joined to all other vertices, and the third distance power of the cycle C_{11} (which is further denoted by T_{11}). Thomassen posed as a problem [16, Problem 3] whether the toroidal 6-critical graphs distinct from T_{11} and the graph obtained by applying Hajos's construction to two copies of K_6 are the only 6-critical graphs that can be embedded on the Klein bottle.

We refute Thomassen's conjecture by exhibiting the list of all nine 6-critical graphs that can be embedded on the Klein bottle (the graphs are depicted in Figure 7.1). The same result was independently established by Chenette et al. [3]. The two proofs are different (and we thus agreed to publish two separate papers): Chenette et al. analyzed 6-critical graphs on the Klein bottle to establish the existence of a vertex w whose reduction yields a reduced graph with small faces only. Our approach is based on a systematic generating of all embeddings of 6-critical graphs on the Klein bottle from the complete graph K_6 and is computer-assisted (unlike the proof of Chenette et al.). We also obtain the list of all nonisomorphic embeddings of 6-critical graphs on the Klein bottle. We believe that our proof can be decomputerized (at the expense of massive case analysis) but we decided not to do so in light of the proof of Chenette et al. which is significantly shorter than our decomputerized proof would be.

As we have mentioned, our proof is computer-assisted. In this paper, we outline the main concepts we use and explain the procedure used to generate all embeddings of 6-critical graphs on the Klein bottle. In order to verify the correctness of our programs, we have separately prepared two different programs implementing our procedures and compared their outputs. Further details of the implementation and the source code of one of our programs can be found at http://kam.mff.cuni.cz/~bernard/klein. In this paper, we establish the correctness of used algorithms and refer the reader to the web page for details on implementation. The outcome of our programs is summarized in section 7 where we also briefly discuss the algorithmic corollaries of our results.

2. 6-critical graphs. In this section, we observe basic properties of 6-critical graphs on the Klein bottle. Euler's formula implies that the average degree of a graph embedded on the Klein bottle is at most six. As Sasanuma [15] established that every 6-regular graph that can be embedded on the Klein bottle is 5-colorable, we have the following proposition (observe that no 6-critical graph contains a vertex of degree four or less).

PROPOSITION 2.1. *The minimum degree of every* 6*-critical graph on the Klein bottle is five.*

Let G be a 6-critical graph on the Klein bottle and v a vertex of degree five in G. Further let v_i, $1 \le i \le 5$, be the neighbors of v in G. If all vertices v_i and v_j, $1 \le i < j \le 5$, are adjacent, the vertices v and v_i, $1 \le i \le 5$, form a clique of order six in G. As G is 6-critical, G must then be a complete graph of order six. Hence, we can conclude the following.

PROPOSITION 2.2. *Let G be a* 6*-critical graph embedded on the Klein bottle. If G is not a complete graph of order six, then G contains a vertex v of degree five that has two nonadjacent neighbors v' and v''.*

We now introduce the following *reduction*: let G be a 6-critical graph embedded

on the Klein bottle that is not isomorphic to K_6 and let v, v', and v'' be three vertices as in Proposition 2.2. $G|vv'v''$ is the graph obtained from G by removing all the edges incident with v except for vv' and vv'' and contracting the edges vv' and vv'' to a new vertex w. The obtained graph can have parallel edges, but it does not have loops as the vertices v' and v'' are not adjacent. Observe that the graph $G|vv'v''$ is not 5-colorable: otherwise, consider a 5-coloring of $G|vv'v''$ and color the vertices v' and v'' with the color assigned to the vertex w. Next, extend the 5-coloring to v—this is possible since v has five neighbors and at least two of them (v' and v'') have the same color. Hence, we obtain a 5-coloring of G.

Since $G|vv'v''$ has no 5-coloring, it contains a 6-critical subgraph—this subgraph will be denoted by $|G|vv'v''|$, and we say that G can be *reduced* to $|G|vv'v''|$. Observe that the reduction operation can again be applied to $|G|vv'v''|$ until a graph that is isomorphic to K_6 is obtained.

We continue with a simple observation on the graph $|G|vv'v''|$.

PROPOSITION 2.3. *Let G be a 6-critical graph embedded on the Klein bottle, v a vertex of degree five in G, and v' and v'' two nonadjacent neighbors of v. The graph $|G|vv'v''|$ contains the vertex w obtained by contracting the path $v'vv''$. Moreover, the vertex w has a neighbor w' in $|G|vv'v''|$ that is a neighbor of v' in G but not of v'' and it also has a neighbor w'' that is a neighbor of v'' but not of v' in G.*

Proof. If $|G|vv'v''|$ does not contain the vertex w, then $|G|vv'v''|$ is a subgraph of $G \setminus \{v, v', v''\}$. Since both $|G|vv'v''|$ and G are 6-critical graphs, this is impossible. Hence, $|G|vv'v''|$ contains the vertex w.

Assume now that $|G|vv'v''|$ contains no vertex w' as described in the statement of the proposition; i.e., all neighbors of w in $|G|vv'v''|$ are neighbors of v'' in G. This implies that $|G|vv'v''|$ is isomorphic to a subgraph of $G \setminus \{v, v'\}$ (view the vertex v'' as w), which is impossible since both G and $|G|vv'v''|$ are 6-critical. A symmetric argument yields the existence of a vertex w''. □

The strategy of our proof is to generate all 6-critical graphs by reversing the reduction operation. More precisely, we choose a vertex w of a 6-critical graph G and partition the neighbors of w into two nonempty sets W_1 and W_2. We next replace the vertex w with a path w_1ww_2 and join the vertex w_i, $i = 1, 2$, to all vertices in the set W_i. Let $G[w, W_1, W_2]$ be the resulting graph. We say that $G[w, W_1, W_2]$ is obtained by *expanding* the graph G. By Proposition 2.3, the following proposition holds (choose w as in the statement of the proposition).

PROPOSITION 2.4. *Let G be a 6-critical graph embedded on the Klein bottle and let v be a vertex of degree five of G with two nonadjacent neighbors v' and v''. The graph $G' = |G|vv'v''|$ contains a vertex w such that $G'[w, W_1, W_2] \subseteq G$ for some partition W_1 and W_2 of the neighbors of the vertex w.*

3. Minimal graphs. Our plan is to generate all 6-critical graphs from the complete graph K_6 by expansions and insertions of new graphs into faces. In this section, we describe the graphs we have to insert into the faces to be sure that we have generated all 6-critical graphs.

A plane graph G with the outer face bounded by a cycle C of length k is said to be *k-minimal* if for every edge $e \in E(G) \setminus C$, there exists a proper precoloring φ_e of C with five colors that cannot be extended to G and that can be extended to a proper 5-coloring of $G \setminus e$ (the graph G with the edge e removed). Note that the precolorings φ_e can differ for various choices of e.

The cycle C_k of length k is k-minimal (the definition vacuously holds); we say that C_k is a *trivial* k-minimal graph. For $k = 3$, it is easy to observe that C_3 is

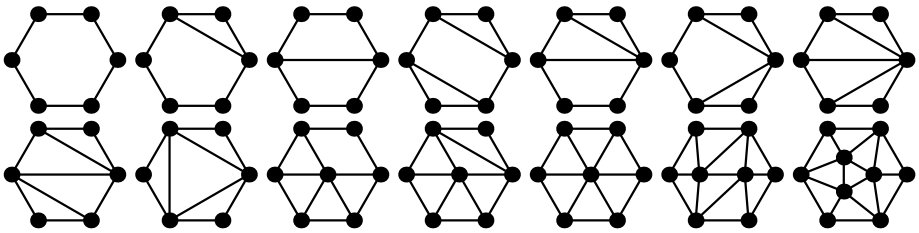


FIG. 3.1. *The list of all* 6*-minimal graphs.*

the only 3-minimal graph since the colors of the vertices of C_3 must differ and every planar graph is 5-colorable. Similarly, every precoloring of a chordless C_4 can be extended to a 5-coloring of its interior and and thus C_4 and C_4 with a chord are the only 4-minimal graphs. As for $k = 5$, Thomassen [16] showed that if G is a plane graph with the outer face bounded by a cycle C of length five and C is chordless, then a precoloring of C with five colors can be extended to G unless G is the 5-wheel and the vertices of C are precolored with all five colors. Hence, C_5, C_5 with one chord, C_5 with two chords, and the 5-wheel are the only 5-minimal graphs. The analogous classification result of Thomassen [16] implies that the only 6-minimal graphs (up to an isomorphism) are those depicted in Figure 3.1; see also [4].

The following lemma justifies the use of k-minimal graphs in our considerations.

LEMMA 3.1. *Let G be a 6-critical graph embedded on the Klein bottle. If C is a contractible cycle of G of length k, then the subgraph G' of G inside the cycle C (G' includes the cycle C itself) is k-minimal.*

Proof. We verify that G' is k-minimal. Let e be an edge of G' that is not contained in C. Let φ_e be the 5-coloring of $G \setminus e$ restricted to the cycle C. Clearly, φ_e cannot be extended to G' but can be extended to $G' \setminus e$. □

In the light of Lemma 3.1, our next goal is to find all k-minimal graphs for small values of k. The following proposition enables us to systematically generate all k-minimal graphs for any fixed k from the lists of k'-minimal graphs for $3 \leq k' < k$.

PROPOSITION 3.2. *If G is a nontrivial k-minimal graph, $k \geq 3$, with the outer cycle C, then either the cycle C contains a chord or G contains a vertex v adjacent to at least three vertices of the cycle C. In addition, if C' is a cycle of G of length k' and G' is the subgraph of G bounded by the cycle C' (inclusively), then G' is a k'-minimal graph.*

Proof. First assume that C is chordless and each vertex v of G is adjacent to at most two vertices of C. Let G' be the subgraph of G induced by the vertices not lying on C. We consider the following list coloring problem: each vertex of G' not incident with the outer face receives a list of all five available colors, and each vertex incident with the outer face is given a list of the colors distinct from the colors assigned to its neighbors on C in G. By our assumption, each such vertex of G' has a list of at least three colors. A classical list coloring result of Thomassen [17] on list 5-colorings of planar graphs yields that G' has a coloring from the constructed lists. Hence, every precoloring of the boundary of G can be extended to the whole graph G, and thus G cannot be k-minimal. This establishes the first part of the proposition. The proof of the fact that every cycle of length k' bounds a k'-minimal subgraph is very analogous to that of Lemma 3.1 and is omitted. □

Proposition 3.2 suggests the following algorithm for generating k-minimal graphs. Assume that we have already generated all ℓ-minimal graphs for $\ell < k$ and let M_ℓ

TABLE 3.1
The numbers of nonisomorphic k-minimal graphs for $3 \leq k \leq 10$ and the largest number n_k of internal vertices of a k-minimal graph.

k	3	4	5	6	7	8	9	10
$\lvert M_k \rvert$	1	2	4	14	46	291	2124	19876
n_k	0	0	1	3	4	6	7	9

be the list of all ℓ-minimal graphs. Note that we have explicitly described the lists M_3, M_4, M_5, and M_6. The list M_k is then generated by the following procedure (the vertices of outer boundary are denoted by $v_1, \ldots, v_k$):

```
M_k := { the cycle C_k on v_1,...,v_k }
repeat
  M' := M_k
  forall 1 <= a < b <= k with b-a>=2 do
    G := the cycle C_k on v_1,...,v_k with the chord v_av_b
    forall G_1 in M_{b-a+1} and G_2 in M_{k+a-b+1} do
      H := G with G_1 and G_2 pasted into its faces
      if H is k-minimal and H is not in M_k then
        add H to M_k
    endfor
  endfor
  forall 1 <= a < b < c <= k do
    G := the cycle C_k on v_1,...,v_k with the vertex v
         adjacent to v_a, v_b and v_c
    forall G_1 in M_{b-a+2}, G_2 in M_{c-b+2} and
           G_3 in M_{k+a-c+2} do
      H := G with G_1, G_2 and G_3 pasted into its faces
      if H is k-minimal and H is not in M_k then
        add H to M_k
    endfor
  endfor
until M_k = M'
```

Proposition 3.2 implies that the list M_k contains all k-minimal graphs after the termination of the procedure: if G is a k-minimal graph, it contains either a chord or a vertex v adjacent to three vertices on the outer cycle and the graphs inside the faces of the skeleton formed by the outer cycle and the chord/edges adjacent to v are also minimal. The verifications of whether the graph G is isomorphic to one of the graphs in M_k and whether G is k-minimal are straightforward, and the reader can find the details in the program available at http://kam.mff.cuni.cz/~bernard/klein.

The numbers of nonisomorphic k-minimal graphs for $3 \leq k \leq 10$ can be found in Table 3.1. We finish this section by justifying our approach with showing that the number of k-minimal graphs is finite for every k; in particular, the procedure always terminates.

PROPOSITION 3.3. *The number of k-minimal graphs is finite for every $k \geq 3$.*

Proof. Let A_k be the number of k-minimal graphs and $A_{k,\ell}$ the number of k-minimal graphs G such that exactly ℓ precolorings of the boundary of G with five colors can be extended to G. Clearly, $A_{k,\ell} = 0$ for $\ell > 5 \cdot 4^{k-1}$ since there are at most $5 \cdot 4^{k-1}$ proper precolorings of the boundary of G. We prove that the numbers $A_{k,\ell}$ are finite by the induction on $5^k + \ell$. More precisely, we establish the following

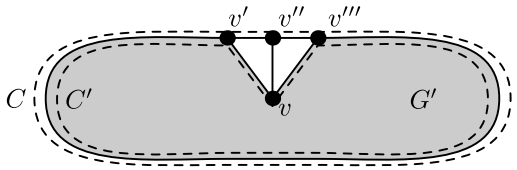


FIG. 3.2. *The notation used in the proof of Proposition* 3.2.

formula:

$$A_{k,\ell} \leq k \cdot \sum_{i=3}^{k-1} 4i(k+2-i)A_i A_{k+2-i} \tag{3.1}$$

$$+ k \cdot \sum_{i=4}^{k-1} \sum_{i'=4}^{k+3-i} 8ii'(k+6-i-i')A_i A_{i'} A_{k+6-i-i'} \tag{3.2}$$

$$+ k \sum_{i=1}^{\ell-1} 2kA_{k,i}. \tag{3.3}$$

Fix k and ℓ. By Proposition 3.2, every k-minimal graph G with ℓ extendable precolorings of its boundary cycle C contains either a chord or a vertex v adjacent to three vertices on C. In the former case, the cycle C and the chord form cycles of length i and $k+2-i$. Since these cycles bound i-minimal and $(k+2-i)$-minimal graphs by Proposition 3.2, the number of such k-minimal graphs is at most $A_i A_{k+2-i}$. After considering at most k possible choices of the chord (for fixed i) and $2i$ and $2(k+2-i)$ possible rotations and/or reflections, we obtain the term (3.1).

Let us analyze the case that G contains a vertex v adjacent to three vertices on C. If the neighbors of v are not three consecutive vertices of C, then the edges between v and its neighbors delimit cycles of lengths $i \geq 4$, $i' \geq 4$, and $k+6-i-i'$. These cycles bound i-minimal, i'-minimal, and $(k+6-i-i')$-minimal graphs, and their number (including different rotations and reflections) is estimated by the term (3.2).

Assume that the neighbors of v on C are consecutive. Let v', v'', and v''' be the neighbors of v and let G' be the subgraph of G inside the cycle C', where C' is the cycle C with the path $v'v''v'''$ replaced with the path $v'vv'''$ (see Figure 3.2). Fix a precoloring φ_0 of the vertices of C except for v''. Let α be the number of ways in which φ_0 can be extended to v that can also be extended to G'. Similarly, α' is the number of ways in which φ_0 can be extended to v'' that can also be extended to G.

We show that $\alpha \leq \alpha'$. If $\alpha = 0$, then $\alpha' = 0$. If $\alpha = 1$, then $\alpha' > 1$. Finally, if $\alpha > 1$, then $\alpha \leq \alpha'$ as any extension of φ_0 to C also extends to G (note that α' is 3 or 4 depending on $\varphi_0(v')$ and $\varphi_0(v''')$). We conclude that the number of precolorings of C' that can be extended to G' does not exceed the number of precolorings of C extendable to G.

Let φ be the precoloring of C that cannot be extended to G but that can be extended to $G \setminus vv''$ and let φ_0 be the restriction of φ to $C \setminus v''$. It is easy to infer that the value of α for this particular precoloring φ_0 must be equal to one. Hence, the number of precolorings of C' that can be extended to G' is strictly smaller than the number of precolorings of C that can be extended to G. Since G' is a k-minimal graph with fewer precolorings of the boundary that can be extended to G' than the

number of precolorings of C extendable to G, the number of k-minimal graphs G with a vertex v with three consecutive neighbors on C including their possible rotations and reflections is estimated by (3.3). This finishes the proof of the inequality and thus the proof of the whole proposition. □

4. Embeddings of K_6 on the Klein bottle. Subsequent applications of our reduction procedure to a 6-critical graph on the Klein bottle eventually lead to an embedding of the complete graph K_6. The resulting embedding of K_6 is either a 2-cell embedding or not. Recall that an embedding is 2-*cell* if every face is homeomorphic to a disc.

If the resulting embedding of K_6 is not 2-cell, the embedding must be isomorphic to the embedding obtained from the unique embedding of K_6 in the projective plane by inserting a cross-cap into one of its faces. Otherwise, the embedding is isomorphic to one of the seven embeddings of K_6 depicted in Figure 4.1. All 2-cell embeddings of K_6 on the Klein bottle can be easily generated by a simple program that ranges through all 2-cell embeddings of K_6 on surfaces: for each vertex v of K_6, the program generates all cyclic permutations of the other vertices (corresponding to the order in which the vertices appear around v) and chooses which edges alter the orientation. Each such pair of cyclic permutations and alterations of orientations determines uniquely both the embedding and the surface. It is straightforward to compute the genus of the surface and test whether the constructed embedding is not isomorphic to one of the previously found embeddings. The source code of the program can be found at http://kam.mff.cuni.cz/~bernard/klein.

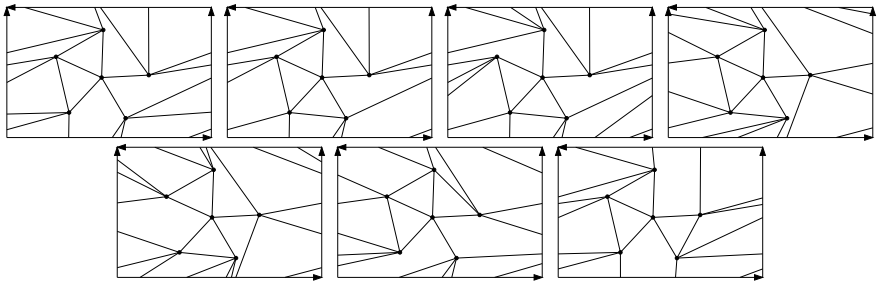


FIG. 4.1. *The list of all seven nonisomorphic* 2*-cell embeddings of* K_6 *on the Klein bottle.*

5. Expansions of 2-cell embeddings of K_6. In this section, we focus on embeddings of 6-critical graphs that can be reduced to a 2-cell embedding of K_6. All such 6-critical graphs can easily be generated, using the expansion operation and Lemma 3.1, by the following procedure:

```
G_1, G_2, G_3, G_4, G_5, G_6, G_7 :=
  non-isomorphic embeddings of K_6 on the Klein bottle
k := 7
i := 1
while i <= k do
  for all vertices w of G_i do
    for all partitions of N(w) into W_1 and W_2 do
       H_0 := G[w,W_1,W_2]
       for all H obtained from H_0 by pasting
                 minimal graphs into its faces do
```

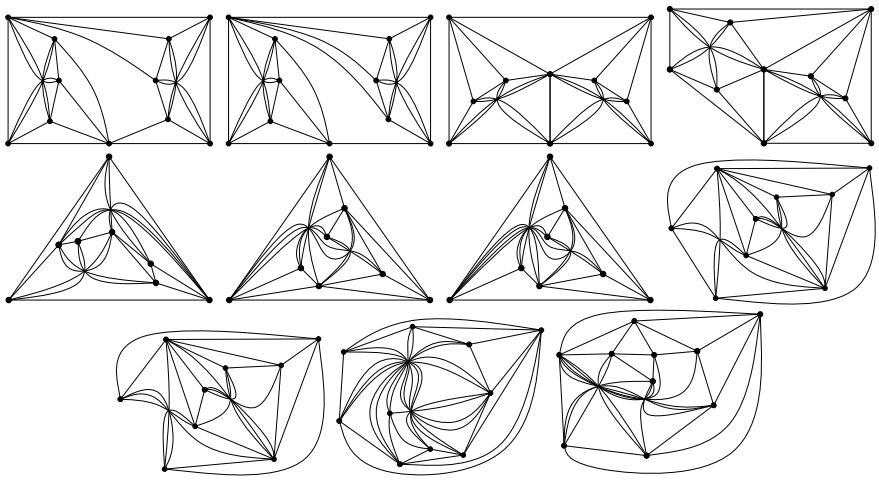


FIG. 5.1. *The list of* 11 *nonisomorphic embeddings of* 6*-critical graphs on the Klein bottle that are distinct from* K_6*. The graphs are drawn in the plane with two cross-caps.*

```
          if H is not isomorphic to any of G_1,..., G_k then
            k := k + 1; G_k := H
          endfor
        endfor
      endfor
    i := i + 1
    done { while }
  output G_1,..., G_k
```

The source code of the program implementing the above procedure can be found at http://kam.mff.cuni.cz/~bernard/klein. The program eventually terminates outputting 11 embeddings of 6-critical graphs on the Klein bottle, which are depicted in Figure 5.1, in addition to the seven 2-cell embeddings of K_6. Hence, Proposition 2.4 and Lemma 3.1 now yield the following lemma.

LEMMA 5.1. *Let* G *be an embedding of a* 6*-critical graph on the Klein bottle that is distinct from* K_6*. If* G *can be sequentially reduced to a* 2*-cell embedding of* K_6 *on the Klein bottle, then* G *is isomorphic to one of the* 11 *embeddings depicted in Figure* 5.1.

6. Expansions of non–2-cell embedding of K_6. As we have already analyzed embeddings of 6-critical graphs that can be reduced to a 2-cell embedding of K_6 on the Klein bottle, it remains to analyze 6-critical graphs that can be reduced to a non–2-cell embedding of K_6. We eventually show that all such embeddings are isomorphic to one of those depicted in Figure 5.1.

LEMMA 6.1. *Let* G *be a* 6*-critical graph embedded on the Klein bottle. If* G *can be reduced to a non–*2*-cell embedding of* K_6*, then* G *is isomorphic to one of the embeddings depicted in Figure* 5.1.

Proof. Let G be a 6-critical graph on the Klein bottle with the smallest order that can be reduced to a non–2-cell embedding of K_6 and that is not isomorphic to any of the embeddings in Figure 5.1. Observe that any possible reduction of G yields a non–2-cell embedding of K_6 on the Klein bottle (otherwise, the reduced graph is a smaller graph missing in Figure 5.1).

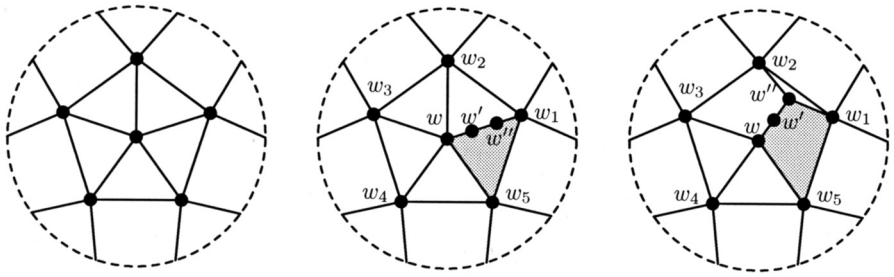


FIG. 6.1. *The unique embedding of K_6 in the projective plane and its two possible expansions.*

Let H be the unique embedding of K_6 in the projective plane and w a vertex of H. By Proposition 2.4, G contains $H[w, W_1, W_2]$ for some partition of the neighborhood of w into nonempty sets W_1 and W_2. By symmetry, $|W_1| = 1$ or $|W_1| = 2$. We first analyze the case that $|W_1| = 1$, i.e., G contains the embedding drawn in the middle of Figure 6.1 as a subgraph. The face which is not 2-cell is drawn using the gray color (the choice is unique since a 6-critical graph cannot contain a separating triangle).

Let G_{15} be the subgraph of G contained inside the cycle $C_{15} = ww'w''w_1w_5$ and let G_{12} be the subgraph contained inside the cycle $C_{12} = ww'w''w_1w_2$. By Lemma 3.1, G_{12} is either the cycle C_{12} with zero, one, or two chords or a 5-wheel bounded by the cycle C_{12}. The interiors of the remaining 2-cell faces of $H[w, W_1, W_2]$ must be empty (since they are triangles).

Assume that $G_{12} \neq C_{12}$. The graph G without the interior of the cycle C_{12} is 5-colorable since G is 6-critical. Observe that the vertices w and w_1 must get the same color in any such 5-coloring (since adding an edge ww_1 to G would form a clique of order six). However, it is always possible to permute the colors of the vertices of G_{15} preserving the colors of w, w_1, and w_5 in such a way that the 5-coloring can be extended to G_{12}. Hence, $G_{12} = C_{12}$.

Since G is 6-critical, the graph G_{15} is 5-colorable. Moreover, the vertices w and w_1 receive distinct colors in every 5-coloring of G_{15}: if the vertices w and w_1 have the same color, the 5-coloring of G_0 can be extended to G.

Let G' be the graph obtained from G_{15} by identifying the vertices w and w_1. Since G_{15} can be drawn in the projective plane with the cycle C_{15} bounding a face, G' can also be drawn in the projective plane. As no 5-coloring assigns the vertices w and w_1 the same color, G' contains K_6 as a subgraph. Since G does not contain K_6 as a subgraph, the subgraph of G' isomorphic to K_6 contains the vertex obtained by the identification of w and w_1. In addition, G' does not contain any edges except for the edges of the complete graph and the path ww_5w_1 (removing any additional edge from G would yield a graph that is also not 5-colorable contrary to our assumption that G is 6-critical). We conclude that G_{15} is composed of

1. the path ww_5w_1, a complete graph on a 5-vertex set X such that $\{w', w''\} \subset X$ and $w_5 \notin X$, and such that $N(w)$ and $N(w_1)$ partition X, or
2. the path ww_5w_1, a complete graph on a 5-vertex set X, $\{w', w'', w_5\} \subset X$, such that $N(w) \setminus \{w_5\}$ and $N(w_1) \setminus \{w_5\}$ partition $X \setminus \{w_5\}$.

In the former case, the graph G is isomorphic to the first or the second embedding in the first line in Figure 5.1; in the latter case, G is isomorphic to the third or the fourth embedding in the first line in the figure.

We now assume that $|W_1| = 2$. $G[w, W_1, W_2]$ is depicted in the right part of

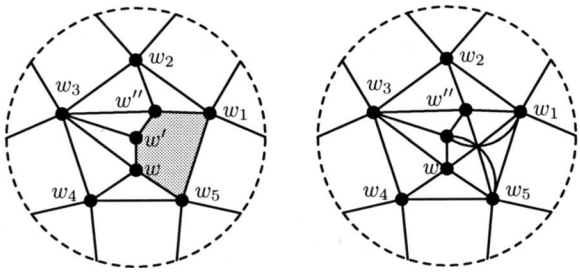


FIG. 6.2. *The embeddings obtained in the analysis in the proof of Lemma* 6.1.

Figure 6.1. We can also assume that w is not adjacent to w_2 in G since otherwise we could choose $W_1 = \{w_1\}$ which would bring us to the previous case. Similarly, the vertices w, w_1, and w_5 do not form a triangular face of G. Let C_{15} be the cycle $ww'w''w_1w_5$, C_{23} the cycle $ww'w''w_2w_3$, G_{15} the subgraph of G inside C_{15}, and G_{23} the subgraph inside C_{23}. Again, G_{23} is either the cycle C_{23} with zero, one, or two chords or a 5-wheel bounded by C_{23}.

It is straightforward (but tedious) to check that any coloring c of G_{15} with five colors extends to a coloring of G unless

- the vertices w and w'' are assigned the same color in c, or
- all the five vertices w, w', w'', w_1, and w_5 are assigned mutually distinct colors and G contains edges w_3w' and w_3w'' (see the embedding in the left part of Figure 6.2).

The reader is asked to verify the details.

We first show that there is a coloring of G_{15} of the latter type. Let G' be the graph obtained from G_{15} by adding the edge ww''. Assume that G' contains a complete graph of order six as a subgraph. If G_{23} contains an inner edge e, consider a 5-coloring of $G \setminus e$ which exists since G is 6-critical. The coloring must assign the vertices w and w'' the same color (since otherwise, c restricted to G_{15} would also be a proper coloring of G'). Consequently, none of the vertices w_i, $1 \leq i \leq 5$, can be assigned the common color of w and w'', which is impossible since the vertices w_i, $1 \leq i \leq 5$, form a clique. We conclude that G_{23} is formed by the cycle C_{23} only. As in the previous case, we can now establish that G' is formed by a subgraph isomorphic to K_6 and the path ww_5w_1w'' and that the vertex w' is contained in the subgraph isomorphic to K_6. This embedding is isomorphic to the first or the last embedding in the first line of Figure 5.1.

Since G' does not contain K_6 as a subgraph, there is a coloring of G_{15} with five colors that assigns w and w'' distinct colors. Hence, G_{15} has a coloring assigning all the vertices w, w', w'', w_1, and w_5 distinct colors and G must be of the type depicted in the left part of Figure 6.2. Since the vertices w and w_2 are not adjacent in G and the degree of w_4 is five, we can consider the graph $|G|w_4ww_2|$; let G_0 be this graph. By the choice of G, G_0 is a non–2-cell embedding of K_6 in the projective plane, and Proposition 2.3 implies that G_0 contains the vertex w_0 obtained by contracting the path ww_4w_2 in G.

If G_0 does not contain the vertex w_3, consider a coloring of G_{15} assigning the vertices w, w', w'', w_1, and w_5 five distinct colors. This coloring restricted to G_0 is a proper coloring of $G_0 = K_6$ with five colors since G_0 can contain only the edges w_0w'' and w_0w_1 in addition to those contained in G_{15} (considering the vertices w and w_0 to be the same vertex). Hence, G_0 contains the vertex w_3. Since the only neighbors of

w_3 in $G|w_4ww_2$ are the vertices w_0, w', w'', w_1, and w_5, the vertex set of G_0 must be $\{w_0, w', w'', w_1, w_5, w_3\}$. In particular, the vertex w_5 is adjacent to w' and w'' in G. A symmetric argument applied to $|G|w_2w''w_4|$ implies that w_1 is adjacent to w' and w'' in G. This brings us to the embedding depicted in the right part of Figure 6.2, which is isomorphic to the third embedding in the second line in Figure 5.1. □

7. Main result. We now summarize the results obtained in the previous sections. The discussion in section 4 and Lemmas 5.1 and 6.1 yield the following theorem.

THEOREM 7.1. *There are nine nonisomorphic* 6*-critical graphs that can be embedded on the Klein bottle which are depicted in Figure* 7.1*. The graphs have altogether a single non–*2*-cell embedding and* 18 *nonisomorphic* 2*-cell embeddings on the Klein bottle, which are depicted in Figures* 4.1 *and* 5.1.

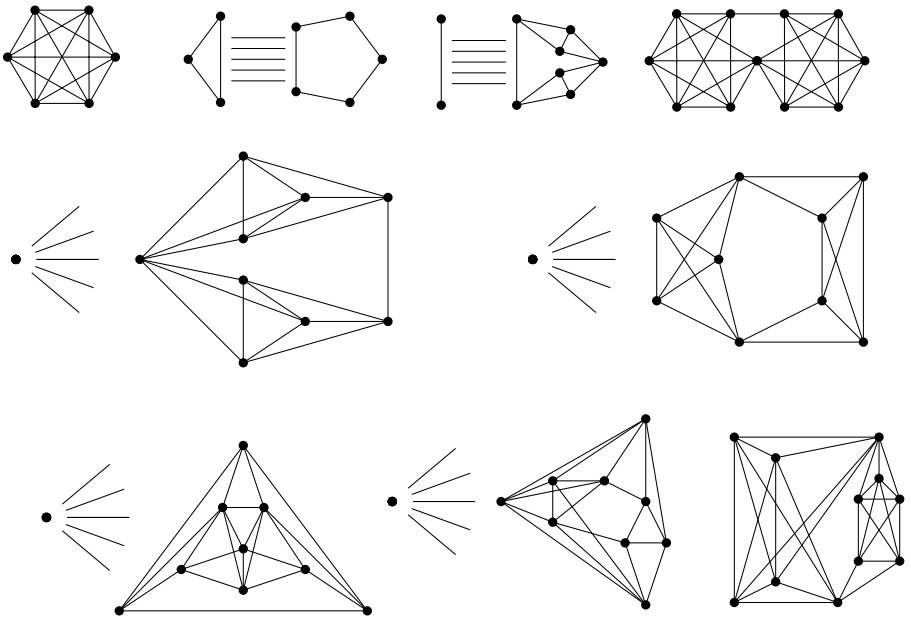


FIG. 7.1. *The list of all nine* 6*-critical graphs that can be embedded in the Klein bottle. Some of the edges are only indicated in the figure: the straight edges between two parts represent that the graph is obtained as the join of the two parts and the vertices with "stars" of edges are adjacent to all vertices in the graph.*

Immediate corollaries of Theorem 7.1 are the following.

COROLLARY 7.2. *Let G be a graph that can be embedded on the Klein bottle. G is* 5*-colorable unless it contains one of the nine graphs depicted in Figure* 7.1 *as a subgraph.*

COROLLARY 7.3. *Let G be a graph embedded on the Klein bottle. G is* 5*-colorable unless it contains a subgraph with embedding isomorphic to one of the* 19 *embeddings depicted in Figures* 4.1, 5.1, *and* 6.1.

Eppstein [7, 8] showed that testing the existence of a subgraph isomorphic to a fixed graph H of a graph embedded on a fixed surface can be solved in linear time. As we have found the explicit list of 6-critical graphs on the Klein bottle, we also obtain the following corollary.

COROLLARY 7.4. *There is an explicit linear-time algorithm for testing whether a graph embedded on the Klein bottle is* 5*-colorable.*

REFERENCES

[1] K. Appel and W. Haken, *Every planar map is four colorable, Part* I. *Discharging*, Illinois J. Math., 21 (1977), pp. 429–490.
[2] K. Appel, W. Haken, and J. Koch, *Every planar map is four colorable, Part* II. *Reducibility*, Illinois J. Math., 21 (1977), pp. 491–567.
[3] N. Chenette, L. Postle, N. Streib, R. Thomas, and C. Yerger, *Five-coloring graphs on the Klein bottle*, submitted.
[4] M. DeVos, K.-I. Kawarabayashi, and B. Mohar, *Locally planar graphs are* 5*-choosable*, J. Combin. Theory Ser. B, 98 (2008), pp. 1215–1232.
[5] G. A. Dirac, *Map colour theorems related to the Heawood colour formula*, J. London Math. Soc., 31 (1956), pp. 460–471.
[6] G. A. Dirac, *A theorem of R. L. Brooks and a conjecture of H. Hadwiger*, Proc. London Math. Soc. (3), 7 (1957), pp. 161–195.
[7] D. Eppstein, *Subgraph isomorphism in planar graphs and related problems*, in Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), San Francisco, CA, 1995, pp. 632–640.
[8] D. Eppstein, *Subgraph isomorphism in planar graphs and related problems*, J. Graph Algorithms Appl., 3 (1999), pp. 1–27.
[9] S. Fisk, *The non-existence of colorings*, J. Combin. Theory Ser. B, 24 (1978), pp. 247–248.
[10] S. Fisk and B. Mohar, *Coloring graphs without short non-bounding cycles*, J. Combin. Theory Ser. B, 60 (1994), pp. 268–276.
[11] T. Gallai, *Kritische Graphen* I, Publ. Math. Inst. Hungar. Acad. Sci., 8 (1963), pp. 165–192.
[12] T. Gallai, *Kritische Graphen* II, Publ. Math. Inst. Hungar. Acad. Sci., 8 (1963), pp. 373–395.
[13] P. J. Heawood, *Map colour theorem*, Quart. J. Pure Appl. Math., 24 (1890), pp. 332–338.
[14] N. Robertson, D. P. Sanders, P. D. Seymour, and R. Thomas, *The four-colour theorem*, J. Combin. Theory Ser. B, 70 (1997), pp. 2–44.
[15] N. Sasanuma, *Chromatic numbers of* 6*-regular graphs on the Klein bottle*, Discrete Math., to appear.
[16] C. Thomassen, *Five-coloring graphs on the torus*, J. Combin. Theory Ser. B, 62 (1994), pp. 11–33.
[17] C. Thomassen, *Every planar graph is* 5*-choosable*, J. Combin. Theory Ser. B, 62 (1994), pp. 180–181.
[18] C. Thomassen, *Color-critical graphs on a fixed surface*, J. Combin. Theory Ser. B, 70 (1997), pp. 67–100.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 384–400

TORIC SURFACE CODES AND MINKOWSKI LENGTH OF POLYGONS*

IVAN SOPRUNOV† AND JENYA SOPRUNOVA‡

To our teacher Askold Khovanskii on the occasion of his 60th anniversary, with love

Abstract. In this paper we prove new lower bounds for the minimum distance of a toric surface code $\mathcal{C}_P$ defined by a convex lattice polygon $P \subset \mathbb{R}^2$. The bounds involve a geometric invariant $L(P)$, called the full Minkowski length of P. We also show how to compute $L(P)$ in polynomial time in the number of lattice points in P.

Key words. evaluation codes, toric codes, Minkowski sum

AMS subject classifications. 94B27, 14G50, 52B20

DOI. 10.1137/080716554

Introduction. Consider a convex polygon P in $\mathbb{R}^2$ whose vertices lie in the integer lattice $\mathbb{Z}^2$. It determines a vector space $\mathcal{L}_K(P)$ (over a field K) of polynomials $f(t_1, t_2)$ whose monomials correspond to the lattice points in P:

$$\mathcal{L}_K(P) = \operatorname{span}_K\{t_1^{m_1} t_2^{m_2} \mid (m_1, m_2) \in P \cap \mathbb{Z}^n\}.$$

Let $\mathbb{F}_q$ be a finite field and $\overline{\mathbb{F}}_q$ its algebraic closure. The *toric surface code* $\mathcal{C}_P$, first introduced by Hansen in [6], is defined by evaluating the polynomials in $\mathcal{L}_{\mathbb{F}_q}(P)$ at all of the points (t_1, t_2) in the algebraic torus $(\mathbb{F}_q^*)^2$. To be more precise, $\mathcal{C}_P$ is a linear code whose codewords are the strings $(f(t_1, t_2) \mid (t_1, t_2) \in (\mathbb{F}_q^*)^2)$ for $f \in \mathcal{L}_{\mathbb{F}_q}(P)$. It is convenient to assume that P is contained in the square $K_q^2 = [0, q-2]^2$ so that all of the monomials in $\mathcal{L}_{\mathbb{F}_q}(P)$ are linearly independent over $\mathbb{F}_q$. Thus $\mathcal{C}_P$ has block length $(q-1)^2$ and the dimension equal to the number of the lattice points in P.

Note that the weight of each nonzero codeword in $\mathcal{C}_P$ is the number of points $(t_1, t_2) \in (\mathbb{F}_q^*)^2$ where the corresponding polynomial does not vanish. Therefore, the minimum distance of $\mathcal{C}_P$ (which is the minimum weight for linear codes) equals

$$d(\mathcal{C}_P) = (q-1)^2 - \max_{0 \neq f \in \mathcal{L}_{\mathbb{F}_q}(P)} Z(f),$$

where $Z(f)$ is the number of zeros (i.e., points of vanishing) in $(\mathbb{F}_q^*)^2$ of f.

The name *toric surface code* comes from the fact that P defines a toric surface X over $\overline{\mathbb{F}}_q$ (strictly speaking the fan that defines X is a refinement of the normal fan of P), where $\mathcal{L}_{\overline{\mathbb{F}}_q}(P)$ can be identified with the space of global sections of a semiample divisor on X (see, for example, [5]). This allows one to exploit algebraic geometric techniques to produce results about the minimum distance of $\mathcal{C}_P$. In particular, Little and Schenck in [10] used intersection theory on toric surfaces to come up with the following general idea: If q is sufficiently large, then polynomials $f \in \mathcal{L}_{\mathbb{F}_q}(P)$ with

*Received by the editors February 28, 2008; accepted for publication (in revised form) September 15, 2008; published electronically January 14, 2009.
http://www.siam.org/journals/sidma/23-1/71655.html
†Department of Mathematics, Cleveland State University, 2121 Euclid Ave., Cleveland, OH 44115 (i.soprunov@csuohio.edu).
‡Department of Mathematical Sciences, Kent State University, Summit Street, Kent, OH 44242 (soprunova@math.kent.edu).

more absolutely irreducible factors will necessarily have more zeros in $(\mathbb{F}_q^*)^2$ (see [10, Proposition 5.2]).

In this paper we expand this idea to produce explicit bounds for the minimum distance of $\mathcal{C}_P$ in terms of certain geometric invariant $L(P)$, which we call the full Minkowski length of P. Essentially $L(P)$ tells you the largest possible number of absolutely irreducible factors a polynomial $f \in \mathcal{L}_{\mathbb{F}_q}(P)$ can have, but it derives it from the geometry of the polygon P (see Definition 1.1). The number $L(P)$ is easily computable—we give a simple algorithm which is polynomial in the number of lattice points in P. Moreover, we obtain a description of the factorization $f = f_1 \cdots f_{L(P)}$ for $f \in \mathcal{L}_{\mathbb{F}_q}(P)$ with the largest number of factors. More precisely, in Proposition 2.3 we show that the Newton polygon P_{f_i} (which is the convex hull of the exponents of the monomials in f_i) is either a primitive segment, a unit simplex, or a triangle with exactly one interior and three boundary lattice points, called an *exceptional triangle*. This description enables us to prove the following bound.

THEOREM 1. *Let $P \subset K_q^2$ be a lattice polygon with area A and full Minkowski length L. Then for $q \geq \max\left(23, \left(c + \sqrt{c^2 + 5/2}\right)^2\right)$, where $c = A/2 - L + 9/4$, the minimum distance of the toric surface code $\mathcal{C}_P$ satisfies*

$$d(\mathcal{C}_P) \geq (q-1)^2 - L(q-1) - \lfloor 2\sqrt{q} \rfloor + 1.$$

The condition that no factorization $f = f_1 \cdots f_{L(P)}$ contains an exceptional triangle (as the Newton polygon of one of the factors) is geometric and can be easily checked for any given P (we provide a simple algorithm for this which is polynomial in the number of lattice points in P). In this case we have a better bound for the minimum distance of the toric surface code.

THEOREM 2. *Let $P \subset K_q^2$ be a lattice polygon with area A and full Minkowski length L. Under the above condition on P, for $q \geq \max\left(37, \left(c + \sqrt{c^2 + 2}\right)^2\right)$, where $c = A/2 - L + 11/4$, the minimum distance of the toric surface code $\mathcal{C}_P$ satisfies*

$$d(\mathcal{C}_P) \geq (q-1)^2 - L(q-1).$$

We remark that our thresholds for q, where the bounds begin to hold, are much smaller than the ones in Little and Schenck's result (see [10, Proposition 5.2]).

Although, as mentioned above, the minimum distance problem for toric codes is tightly connected to toric varieties, our methods are geometric and combinatorial and do not use algebraic geometry, except for the Hasse–Weil bound adapted to toric surfaces (see section 2.2). In section 1 we define the full Minkowski length $L(P)$ and establish combinatorial properties of polygons with $L(P) = 1, 2$. In section 2 we give a proof of Theorems 1 and 2. Section 3 is devoted to the above mentioned algorithms for computing $L(P)$ and determining the presence of an exceptional triangle. Finally, in section 4 we give a detailed analysis of three toric surface codes which illustrates our methods.

1. Full Minkowski length of polytopes.

1.1. Minkowski sum. Let P and Q be convex polytopes in $\mathbb{R}^n$. Their *Minkowski sum* is

$$P + Q = \{p + q \in \mathbb{R}^n \mid p \in P,\ q \in Q\},$$

which is again a convex polytope. Figure 1 shows the Minkowski sum of a triangle and a square.

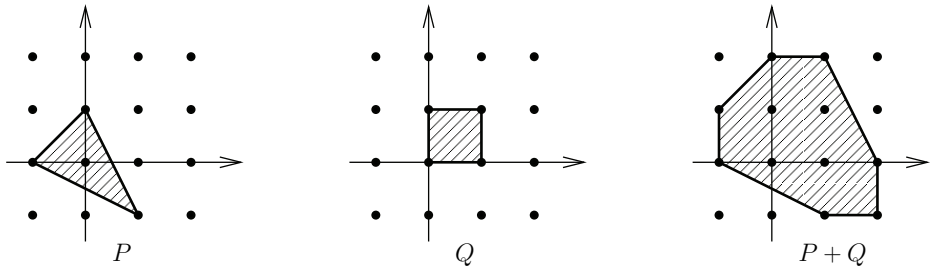


FIG. 1. *The Minkowski sum of two polygons.*

Let f be a Laurent polynomial in $K[t_1^{\pm 1}, \dots, t_n^{\pm 1}]$ (for some field K). Then its *Newton polytope* P_f is the convex hull of the exponent vectors of the monomials appearing in f. Thus P_f is a *lattice polytope* as its vertices belong to the integer lattice $\mathbb{Z}^n \subset \mathbb{R}^n$. Note that if $f, g \in K[t_1^{\pm 1}, \dots, t_n^{\pm 1}]$, then the Newton polytope of their product P_{fg} is the Minkowski sum $P_f + P_g$. A *primitive lattice segment* E is a line segment whose only lattice points are its endpoints. The difference of the endpoints is a vector v_E whose coordinates are relatively prime (v_E is defined up to a sign). A polytope which is the Minkowski sum of primitive lattice segments is called a *(lattice) zonotope*.

The automorphism group of the lattice is the group of affine unimodular transformations, denoted by $\mathrm{AGL}(n, \mathbb{Z})$, which consists of translations by an integer vector and linear transformations in $\mathrm{GL}(n, \mathbb{Z})$. Affine unimodular transformations correspond to monomial changes of variables in $K[t_1^{\pm 1}, \dots, t_n^{\pm 1}]$ and preserve the zero set of f in the algebraic torus $(K^*)^n$.

1.2. Full Minkowski length. Let P be a lattice polytope in $\mathbb{R}^n$. Consider a Minkowski decomposition

$$P = P_1 + \cdots + P_\ell$$

into lattice polytopes P_i of positive dimension. Clearly, there are only finitely many such decompositions. We let $\ell(P)$ be the largest number of summands in such decompositions of P, and call it the *Minkowski length* of P.

DEFINITION 1.1. *The full Minkowski length of P is the maximum of the Minkowski lengths of all subpolytopes Q in P,*

$$L(P) := \max\{\ell(Q) \,|\, Q \subseteq P\}.$$

A subpolytope $Q \subseteq P$ is called maximal for P if $\ell(Q) = L(P)$. A Minkowski decomposition of Q into $L(P)$ summands of positive dimension will be referred to as a maximal (Minkowski) decomposition in P.

Here are a few simple properties of $L(P)$ and maximal subpolytopes.

PROPOSITION 1.2. *Let P, P_1, P_2, and Q be lattice polytopes in $\mathbb{R}^n$.*

(1) *$L(P)$ is $\mathrm{AGL}(n, \mathbb{Z})$-invariant.*
(2) *$L(P) \geq 1$ if and only if $\dim(P) > 0$.*
(3) *If $P_1 + P_2 \subseteq P$, then $L(P_1) + L(P_2) \leq L(P)$.*
(4) *If Q is maximal for P, then Q contains a zonotope Z maximal for P.*

Proof. The first three statements are trivial. For the fourth one, note that if

$$Q = Q_1 + \cdots + Q_{L(P)}$$

is a maximal Minkowski decomposition in P, then by replacing each Q_i with one of its edges we obtain a zonotope $Z \subseteq Q$ with $\ell(Z) \geq L(P)$. But $Z \subseteq P$, so $\ell(Z) = L(P)$. □

Notice that the summands of every maximal decomposition in P are polytopes of full Minkowski length 1. It seems to be a hard problem to describe polytopes of full Minkowski length 1 in general. However, in dimensions 1 and 2 we do have a simple description for such polytopes (Theorem 1.4).

DEFINITION 1.3. *A lattice polytope P is* strongly indecomposable *if its full Minkowski length $L(P)$ is 1. In other words, no subpolytope $Q \subseteq P$ is a Minkowski sum of lattice polytopes of positive dimensions.*

Clearly, primitive segments are strongly indecomposable and are the only one-dimensional strongly indecomposable polytopes.

Let Δ be the standard 2-simplex and T_0 be the triangle with vertices $(1,0)$, $(0,1)$, and $(2,2)$ (see Figure 2). It is easy to see that they are both strongly indecomposable.

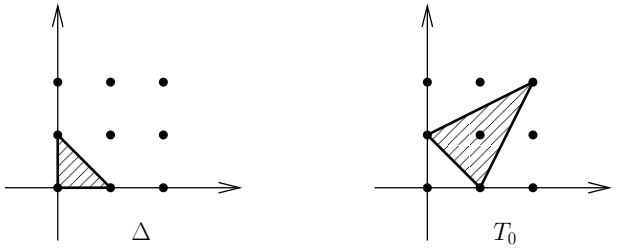


FIG. 2. *Strongly indecomposable polygons.*

The next theorem shows that these are essentially the only strongly indecomposable polygons. In the proof of this theorem and frequently later in the paper we will use Pick's formula: If P is a lattice polygon in $\mathbb{R}^2$, then the area of P equals

$$A = I + \frac{B}{2} - 1,$$

where I is the number of interior lattice points in P and B is the number of boundary points in P. The proof of this formula can be found, for example, in [3].

THEOREM 1.4. *Let P be a strongly indecomposable polygon. Then P is* $\mathrm{AGL}(2,\mathbb{Z})$*-equivalent to either the standard 2-simplex Δ or the triangle T_0 above.*

Proof. First, note that P cannot contain more than four lattice points. Indeed, suppose $a = (a_1, a_2)$ and $b = (b_1, b_2)$ lie in $P \cap \mathbb{Z}^2$. If $a_i \equiv b_i \mod 2$, for $i = 1, 2$, then the segment $[a, b]$ lies in P and is not primitive; hence, $L(P) > 1$. Since there are only four possible pairs of remainders mod 2 and P has at most four lattice points.

Suppose P is a triangle, then its sides must be primitive and either P has no interior lattice points or it has exactly one interior lattice point. In the first case, P has area $1/2$ (by Pick's formula) and so is $\mathrm{AGL}(2,\mathbb{Z})$-equivalent to Δ. In the second case, P has area $3/2$ (by Pick's formula) and hence any two of its sides generate a parallelogram of area 3. Every such triangle is $\mathrm{AGL}(2,\mathbb{Z})$-equivalent to T_0.

Now suppose P is a quadrilateral. Then it has no interior lattice points and so its area is 1 (by Pick's formula). Every such quadrilateral is $\mathrm{AGL}(2,\mathbb{Z})$-equivalent to the unit square. However, the unit square is obviously decomposable. □

DEFINITION 1.5. *A lattice polygon is called a* unit triangle *if it is* $\mathrm{AGL}(2,\mathbb{Z})$*-equivalent to Δ, and an* exceptional triangle *if it is* $\mathrm{AGL}(2,\mathbb{Z})$*-equivalent to T_0.*

The following theorem describes maximal Minkowski decompositions for a given lattice polygon P.

THEOREM 1.6. *Let P be a lattice polygon in $\mathbb{R}^2$ with full Minkowski length $L(P)$. Consider a maximal Minkowski decomposition in P:*

$$Q = Q_1 + \cdots + Q_{L(P)},$$

for some $Q \subseteq P$. Then one of the following holds:

(1) *every Q_i is either a primitive segment or a unit triangle;*
(2) *after an $\mathrm{AGL}(2,\mathbb{Z})$-transformation and reordering of the summands the decomposition is*

$$Q = T_0 + m_1[0,e_1] + m_2[0,e_2] + m_3[0,e_1+e_2],$$

where m_i are nonnegative integers such that $m_1+m_2+m_3 = L(P)-1$ and the e_i are the standard basis vectors.

Proof. Since every Q_i must be strongly indecomposable, by Theorem 1.4 it is a primitive segment, a unit triangle, or an exceptional triangle. We claim that if one of the Q_i is an exceptional triangle, then the other summands are primitive segments in only three possible directions. This follows from the two lemmas below. ∎

LEMMA 1.7. *Consider two primitive segments E_1, E_2 in $\mathbb{Z}^2$, and let v_1, v_2 be the corresponding vectors. If $|\det(v_1,v_2)| \geq 3$, then $L(E_1+E_2) \geq 3$.*

Proof. We can assume that $v_1 = (1,0)$ and $v_2 = (a,b)$ with $0 \leq a < b$ and $b = \det(v_1,v_2)$. Cases when $3 \leq b \leq 6$ are easily checked by hand. For $b \geq 7$ we can use the same argument as in the proof of Theorem 1.4 to show that $\Pi = E_1 + E_2$ contains a segment of lattice length 3. Indeed, the area of Π equals $b \geq 7$. By Pick's formula, Π has at least ten lattice points. But then there exist $a = (a_1,a_2)$ and $b = (b_1,b_2)$ in Π such that $a_i \equiv b_i \bmod 3$, for $i = 1,2$. Therefore the segment $[a,b]$ is contained in Π and has lattice length 3. ∎

LEMMA 1.8. *Let $P \subset \mathbb{R}^2$ be strongly indecomposable. Then $L(T_0+P) \geq 3$ unless P is a primitive segment in the direction of e_1, e_2 or e_1+e_2.*

Proof. Let E_1 be an edge of T_0 and E_2 an edge of P, and let v_1, v_2 be the corresponding vectors. If $|\det(v_1,v_2)| \geq 3$, then by Lemma 1.7 $L(E_1+E_2) \geq 3$, and since $E_1+E_2 \subseteq T_0+P$ we also have $L(T_0+P) \geq 3$. Therefore we suppose that $|\det(v_1,v_2)| \leq 2$ for all edges E_1 in T_0. Then we have the following linear inequalities for $v_2 = (s,t)$:

$$-2 \leq s+t \leq 2, \quad -2 \leq 2s-t \leq 2, \quad -2 \leq s-2t \leq 2.$$

Clearly, the only integer solutions (up to central symmetry) are $v_1 = (1,0)$, $(0,1)$, and $(1,1)$. Now if P contains at least 2 edges in these directions, then it must also contain (up to a translation) either $T = \mathrm{span}\{(0,0),(1,0),(1,1)\}$ or $T = \mathrm{span}\{(0,0),(0,1),(1,1)\}$. But in both cases the sum T_0+T contains a 1×2 rectangle which has Minkowski length three. Therefore, $L(T_0+P) \geq 3$. ∎

Remark 1.9. Notice that in Lemma 1.8 the special directions e_1, e_2 or e_1+e_2 have an easy $\mathrm{AGL}(2,\mathbb{Z})$-invariant description: they are obtained by connecting the interior lattice point in T_0 to the vertices.

While classifying polygons of every given full Minkowski length does not seem feasible, we will make a few statements about polygons of full Minkowski length 2, which we will use later.

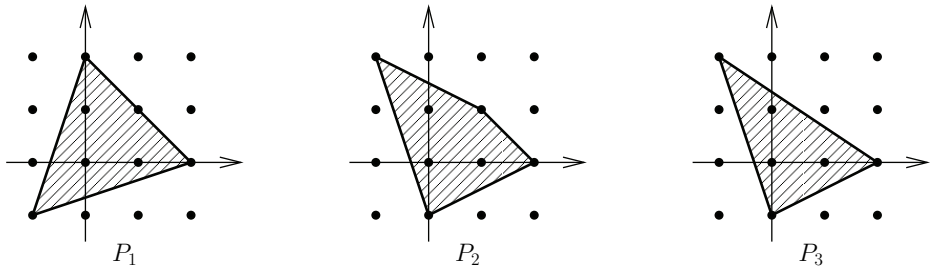


FIG. 3. *Full length two polygons with three interior lattice points.*

PROPOSITION 1.10. *Suppose $L(P) = 2$. Then we have the following:*

(1) *P has at most three interior lattice points, i.e., $I(P) \leq 3$.*
(2) *If $I(P) = 3$, then P is $\mathrm{AGL}(2, \mathbb{Z})$-equivalent to one of the polygons depicted in Figure 3.*
(3) *If $I(P) = 3$, then $L(P + T_0) \geq 4$.*

Proof. (1) The proof is somewhat technical so we will sketch its major steps. Assume P has four or more interior lattice points. First, it is not hard to show that one can choose four interior lattice points in P so that after an $\mathrm{AGL}(2, \mathbb{Z})$-transformation they form either a unit square, $\{(0,0), (1,0), (0,1), (1,1)\}$, or a base 2 isosceles triangle, $\{(-1,0), (0,0), (1,0), (0,1)\}$.

In the first case, note that P must include a lattice point which is distance one from the square and lies on one of the lines containing the sides of the square. By symmetry we can assume it is $(2,0)$. In Figure 4 on the left, the solid dots represent the five points that now belong to P, the crosses represent the points that cannot belong to P (otherwise its length would be greater than 2). Now if point $(0,2)$ does not belong to P (the middle picture in Figure 4), then either $(-1,2)$ or $(1,2)$ does. But in either case the four points of the unit square cannot all lie in the interior of P. If point $(0,2)$ does belong to P, then it produces more forbidden points (the rightmost picture in Figure 4). Then again, it is not hard to see that no such P can exist.

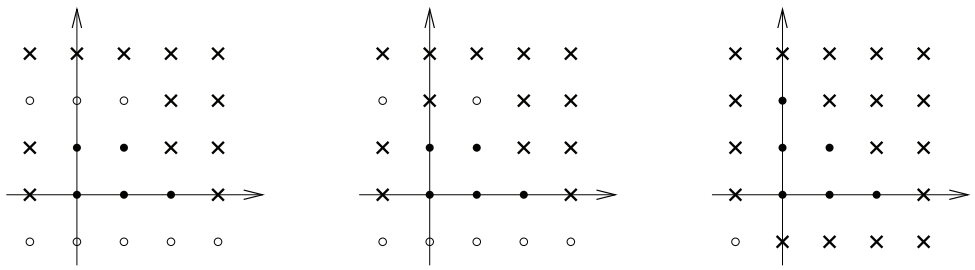


FIG. 4. *Nonexistence of full length two polygons with $I(P) > 3$.*

Playing the same game, one can show that no P exists in the second case as well.

(2) First, one can show that the three interior lattice points cannot be collinear. Thus we can assume that they are $\{(0,0), (1,0), (0,1)\}$. Our first case is when $(1,1)$ also lies in P. Since this must be a boundary point and there are no more interior points in P, we see that $(-1,2)$ and $(0,2)$ are the only possible boundary points of P on the line $y = 2$. Similarly, $(2,0)$ and $(2,-1)$ are the only possible boundary points of P on the line $x = 2$. Since both $(-1,2)$ and $(2,-1)$ cannot belong to P, using symmetry we arrive at two possibilities for the boundary piece of P containing $(1,1)$,

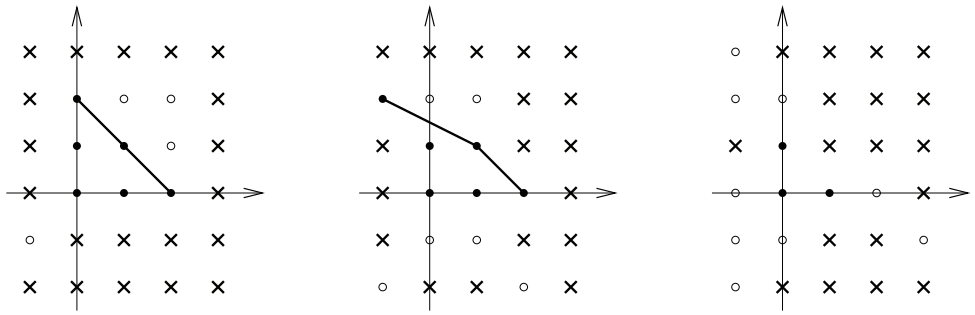


FIG. 5. *Constructing full length two polygons with* $I(P) = 3$.

depicted in Figure 5 on the left. As in part (1), we crossed out the points which cannot appear in P since $L(P) = 2$. Then it becomes clear that the only P (up to symmetry) containing $\{(0,0),(1,0),(0,1)\}$ and $(1,1)$ are P_1 and P_2 in Figure 3.

In the second case, when $(1,1)$ does not lie in P we can assume that $(1,-1)$ and $(-1,1)$ do not lie in P either, otherwise we can reduce it to the previous case by a unimodular transformation. Also, both $(2,-1)$ and $(-1,2)$ cannot lie in P, therefore by symmetry we can assume that $(2,-1)$ does not. As before, by crossing out forbidden points we obtain the rightmost picture in Figure 5. Now it is easy to see that the only P containing the three points in the interior is P_3 in Figure 3.

(3) By (2) it is enough to check that $L(P_i + T) \geq 4$ for every $1 \leq i \leq 3$ and any exceptional triangle T.

We first look at P_1. By Lemma 1.8 and Remark 1.9 we have $L(E+T) \geq 3$ for any primitive segment E except for the three special segments E_1, E_2, E_3 that connect the interior lattice point of T to its vertices. If $T \neq T_0$, then one of $[0,e_1]$, $[0,e_2]$, $[0,e_1+e_2]$ is not among the E_i. But P_1 contains the segments $2[0,e_1], 2[0,e_2]$, and $(-1,-1)+2[0,e_1+e_2]$. If, say, $[0,e_1]$ is not among the E_i, then $L(2[0,e_1]+T) \geq 4$ and hence $L(P_1+T) \geq 4$. It remains to show that $L(P_1+T_0) \geq 4$, which can easily be checked by hand.

A similar argument works for P_3. We only need to replace T_0 with T_0', the triangle with vertices $(0,0),(1,1)$, and $(-1,2)$. Its special segments $[0,e_1]$, $[0,e_2]$, $[0,-e_1+e_2]$ are contained in P with multiplicity 2. Finally, since $P_3 \subset P_2$ we do not need to do any extra work for P_2. ∎

2. Bounds for toric surface codes.

2.1. Toric surface codes. Fix a finite field $\mathbb{F}_q$ where q is prime power. For any convex lattice polygon P in $\mathbb{R}^2$ we associate a $\mathbb{F}_q$-vector space of bivariate polynomials whose monomials have exponent vectors in $P \cap \mathbb{Z}^2$:

$$\mathcal{L}(P) = \operatorname{span}_{\mathbb{F}_q}\{t^m \mid m \in P \cap \mathbb{Z}^2\}, \quad \text{where } t^m = t_1^{m_1} t_2^{m_2}.$$

If P is contained in the square $K_q^2 = [0, q-2]^2$, then the monomials t^m are linearly independent over $\mathbb{F}_q$ and so $\dim \mathcal{L}(P) = |P \cap \mathbb{Z}^2|$. In what follows we will always assume that $P \subset K_q^2$.

The *toric surface code* $\mathcal{C}_P$ is a linear code whose codewords are the strings of values of $f \in \mathcal{L}(P)$ at all points of the algebraic torus $(\mathbb{F}_q^*)^2$ (in some linear order):

$$\mathcal{C}_P = \{\big(f(t), t \in (\mathbb{F}_q^*)^2\big) \mid f \in \mathcal{L}(P)\}.$$

This is a linear code of block length $(q-1)^2$ and dimension $|P\cap\mathbb{Z}^2|$. The weight of each nontrivial codeword equals the number of points $t \in (\mathbb{F}_q^*)^2$ where the corresponding polynomial does not vanish. Let $Z(f)$ denote the number of points in $(\mathbb{F}_q^*)^2$ where f vanishes. Then the minimum distance $d(\mathcal{C}_P)$, which is also the minimum weight, equals

$$d(\mathcal{C}_P) = (q-1)^2 - \max_{0\neq f\in\mathcal{L}(P)} Z(f).$$

2.2. The Hasse–Weil bound. Consider $f \in \mathcal{L}(P)$. Its *Newton polygon* P_f is the convex hull of the lattice points in $\mathbb{R}^2$ corresponding to the monomials in f. We have

$$f(t) = \sum_{m\in P_f\cap\mathbb{Z}^2} \lambda_m t^m, \quad \text{where } t^m = t_1^{m_1}t_2^{m_2}, \quad \lambda_m \in \mathbb{F}_q.$$

Let X be a smooth toric surface over $\overline{\mathbb{F}}_q$ defined by a fan $\Sigma_X \subset \mathbb{R}^2$ which is a refinement of the normal fan of P_f. Then f can be identified with a global section of a semiample divisor on X. Let C_f be the closure in X of the affine curve given by $f = 0$ in $(\overline{\mathbb{F}}_q^*)^2$. If f is absolutely irreducible, i.e., C_f is irreducible, then the number of $\mathbb{F}_q$-rational points $|C_f(\mathbb{F}_q)|$ satisfies the Hasse–Weil bound:

$$|C_f(\mathbb{F}_q)| \le q + 1 + \lfloor 2g\sqrt{q}\rfloor,$$

where g is the *arithmetic genus* of C_f. For the case of smooth curves, see, for example, [11]; for singular curves we refer to [1].

Since we are interested in the number $Z(f)$ of zeros of f in the torus $(\mathbb{F}_q^*)^2$, the above bound might be improved by subtracting possible $\mathbb{F}_q$-rational points on C_f at "infinity." More precisely, we have the following proposition.

PROPOSITION 2.1. *Let f be absolutely irreducible with Newton polygon P_f. Then*

$$Z(f) \le q + 1 + \lfloor 2I(P_f)\sqrt{q}\rfloor - B'(P_f), \tag{2.1}$$

where $I(P_f)$ is the number of interior lattice points and $B'(P_f)$ is the number of primitive edges of P_f.

Proof. It is a classical result from the theory of toric varieties that the arithmetic genus g equals the number of interior lattice points in P_f (see [7] for the general case or [10] for the case of curves).

Let $D \subset X$ be the invariant divisor at "infinity," i.e., $D = X \setminus (\overline{\mathbb{F}}_q^*)^2$. Then the Hasse–Weil bound implies

$$Z(f) \le q + 1 + \lfloor 2I(P_f)\sqrt{q}\rfloor - |C_f(\mathbb{F}_q)\cap D|.$$

The divisor D is the disjoint union of zero- and one-dimensional orbits in X. The one-dimensional orbits are isomorphic to $\overline{\mathbb{F}}_q^*$ and correspond to the rays of Σ_X. Since Σ_X is a refinement of the normal fan of P_f, some of the orbits correspond to the edges of P_f. Let E be an edge of P_f and O_E the corresponding orbit in X, and consider the "restriction" of f to E, i.e., a univariate polynomial $f_E(s)$ whose coefficients are λ_m for $m \in E$, ordered counterclockwise. Then the intersection number $C_f \cdot O_E$ equals the number of zeros of f_E in $\overline{\mathbb{F}}_q^*$ (see [9, Theorem 1 of section 2]). Note that if E is primitive, then f_E is linear, hence, has exactly one $\mathbb{F}_q$-rational zero on O_E. Therefore,

$|C_f(\mathbb{F}_q) \cap D|$ is greater than or equal to $B'(P_f)$, the number of primitive edges of P_f, and the proposition follows. ∎

COROLLARY 2.2. *Let $f \in \mathcal{L}(P)$ be absolutely irreducible and P_f its Newton polygon.*

(1) *If P_f is an exceptional triangle, then $Z(f) \leq q - 2 + \lfloor 2\sqrt{q} \rfloor$.*
(2) *If $I(P_f) = 0$, then $Z(f) \leq q - 1$ unless P_f is twice a unit triangle in which case $Z(f) \leq q + 1$.*

Proof. Part (1) follows immediately from Proposition 2.1. For (2) we use the classification of polygons with no interior lattice points (see, for example, [9] or [2]): P_f is $\mathrm{AGL}(2, \mathbb{Z})$-equivalent to either (a) 2Δ or (b) a trapezoid (see Figure 6) where $0 \leq a \leq b$ (this includes primitive segments when $a = b = 0$ and unit triangles when $a = 0$, $b = 1$). In the first case $Z(f) \leq q + 1$ by (2.1). In the second case P_f has at least two primitive edges, so $Z(f) \leq q - 1$, again by (2.1). ∎

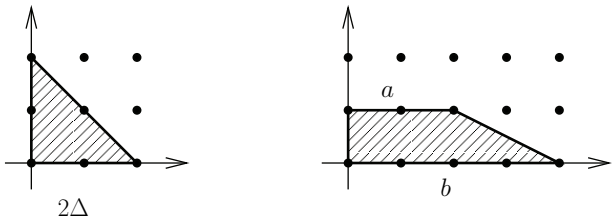


FIG. 6. *Polygons with no interior lattice points.*

2.3. Bounds for the minimum distance. Let $\mathcal{C}_P$ be the toric surface code defined by a lattice polygon P in K_q^2. In this section we prove bounds for the minimum distance of $\mathcal{C}_P$ in terms of the full Minkowski length $L(P)$ of the polygon P.

Here is our first application of the results of the previous section.

PROPOSITION 2.3. *Let $f \in \mathcal{L}(P)$ be a polynomial with the largest number of absolutely irreducible factors, $f = f_1 \cdots f_L$. Then we have the following:*

(1) *$L = L(P)$ and every $P(f_i)$ is either a primitive segment, a unit triangle, or an exceptional triangle.*
(2) *The number of zeros of f in $(\mathbb{F}_q^*)^2$ satisfies*

$$Z(f) \leq L(q-1) + \lfloor 2\sqrt{q} \rfloor - 1.$$

(3) *If $P(f_i)$ is not an exceptional triangle for any $1 \leq i \leq L$, then*

$$Z(f) \leq L(q-1).$$

Proof. Part (1) follows directly from Theorem 1.6. Moreover, the theorem implies that either (a) all P_i are primitive segments or unit triangles or (b) one of the P_i is an exceptional triangle and the others are primitive segments. In the first case every f_i has at most $q - 1$ zeros in $(\mathbb{F}_q^*)^2$ by Corollary 2.2. Not accounting for possible common zeroes of the f_i we obtain the bound in (3). In the second case one of the f_i has at most $q - 2 + \lfloor 2\sqrt{q} \rfloor$ zeros and the others have at most $q - 1$ zeros, again by Corollary 2.2. As before, disregarding possible common zeros of the f_i we get the bound in (2). ∎

The next proposition deals with polynomials f whose number of absolutely irreducible factors is $L(P) - 1$.

PROPOSITION 2.4. *Let P have full Minkowski length L, and let $f \in \mathcal{L}(P)$ have $L-1$ absolutely irreducible factors. Then*

$$Z(f) \leq (L-1)(q-1) + \lfloor 6\sqrt{q} \rfloor.$$

Proof. As before, let $f = f_1 \cdots f_{L-1}$ be the decomposition of f into absolutely irreducible factors, and let P_i be the Newton polygon of f_i. First, by Proposition 1.2

$$k+1 = L \geq \sum_{i=1}^{k} L(P_i) \geq k;$$

hence, up to renumbering, $L(P_1) \leq 2$ and $L(P_i) = 1$ for $2 \leq i \leq k$.

Assume $L(P_1) = 1$. Then every P_i is either a strongly indecomposable triangle or a lattice segment. We claim that at most three of the P_i are exceptional triangles, and so the statement follows from Corollary 2.2. Indeed, if, say, $P_1, \ldots, P_4$ are exceptional triangles, then by Lemma 1.8 $L(P_1 + \cdots + P_4) \geq 6$. Applying Proposition 1.2 again we get

$$k+1 = L \geq L(P_1 + \cdots + P_4) + \sum_{i=5}^{k} L(P_i) \geq 6 + (k-4) = k+2,$$

a contradiction.

Now assume $L(P_1) = 2$. According to (1) in Proposition 1.10, we have $I(P_1) \leq 3$. Also since $L(P_1) = 2$, at most one of the other P_i is an exceptional triangle. This follows from Lemma 1.8 using arguments similar to the previous case. We now have three subcases.

- If $I(P_1) = 1$, then we have

$$Z(f) \leq (q+1+\lfloor 2\sqrt{q} \rfloor) + (q-2+\lfloor 2\sqrt{q} \rfloor) + (L-3)(q-1) \leq (L-1)(q-1) + \lfloor 6\sqrt{q} \rfloor.$$

- If $I(P_1) = 2$, then P_1 has at least one primitive edge which we prove in Lemma 2.5 below. Therefore by Proposition 2.1 we have

$$Z(f) \leq (q+\lfloor 4\sqrt{q} \rfloor) + (q-2+\lfloor 2\sqrt{q} \rfloor) + (L-3)(q-1) \leq (L-1)(q-1) + \lfloor 6\sqrt{q} \rfloor.$$

- Finally, if $I(P_1) = 3$, then none of the other P_i is an exceptional triangle. This follows from Proposition 1.10, (3), and the above arguments. In this case P_1 has at least two primitive edges by Proposition 1.10, (2). Therefore by Proposition 2.1 we have

$$Z(f) \leq (q-1+\lfloor 6\sqrt{q} \rfloor) + (L-2)(q-1) = (L-1)(q-1) + \lfloor 6\sqrt{q} \rfloor. \qquad \square$$

LEMMA 2.5. *If $L(P) = 2$ and $I(P) = 2$, then P has a primitive edge.*

Proof. Since $L(P) = 2$, no edge can have more than 3 lattice points. If P has 4 or more edges, in which none are primitive, then P has at least 8 boundary lattice points and, hence, at least 10 lattice points total. But then P contains a lattice segment of lattice length 3 (see the proof of Lemma 1.7), which contradicts the assumption $L(P) = 2$.

It remains to show that triangles with no primitive edges, 2 interior lattice points, and 6 boundary lattice points do not exist. Let T be such a triangle and let $2E_1$, $2E_2$ be two of its edges, where E_1 and E_2 are primitive. Then E_1, E_2 form a triangle T' of area $A(T') = \frac{1}{4}A(T)$. On the other hand, by Pick's formula $A(P) = 4$, and hence

$A(T') = 1$. This implies that up to an $\mathrm{AGL}(2,\mathbb{Z})$-transformation $E_1 = [0, e_1]$ and $E_2 = [0, e_1 + 2e_2]$, but then $I(T) = 1$, a contradiction. □

Now we are ready for the main result of this section.

THEOREM 2.6. *Let $P \subset K^2_{q-1}$ be a lattice polygon with area $A = A(P)$ and full Minkowski length $L = L(P)$. Then*

(1) *for $q \geq \max\big(23, \big(c+\sqrt{c^2+5/2}\big)^2\big)$, where $c = A/2-L+9/4$, every polynomial $f \in \mathcal{L}(P)$ has at most $L(q-1)+\lfloor 2\sqrt{q}\rfloor - 1$ zeros in $(\mathbb{F}_q^*)^2$. Consequently, the minimum distance for the toric surface code $\mathcal{C}_P$ satisfies*

$$d(\mathcal{C}_P) \geq (q-1)^2 - L(q-1) - \lfloor 2\sqrt{q}\rfloor + 1.$$

(2) *If no maximal decomposition in P contains an exceptional triangle, then for $q \geq \max\big(37, \big(c+\sqrt{c^2+2}\big)^2\big)$, where $c = A/2 - L + 11/4$, every polynomial $f \in \mathcal{L}(P)$ has at most $L(q-1)$ zeros in $(\mathbb{F}_q^*)^2$. Consequently, the minimum distance for the toric surface code $\mathcal{C}_P$ satisfies*

$$d(\mathcal{C}_P) \geq (q-1)^2 - L(q-1).$$

Proof. (1) As we have seen in Proposition 2.3, (2), the bound holds for the polynomials with the largest number of irreducible factors. We are going to show that for large enough q every polynomial with fewer irreducible factors will have no greater than $L(q-1)+\lfloor 2\sqrt{q}\rfloor - 1$ zeros in $(\mathbb{F}_q^*)^2$.

Let $f \in \mathcal{L}(P)$ have $k < L$ absolutely irreducible factors $f = f_1 \cdots f_k$, and let P_i be the Newton polygon of f_i. If $k = L-1$, then we can use the bound in Proposition 2.4:

$$Z(f) \leq (L-1)(q-1) + \lfloor 6\sqrt{q}\rfloor. \tag{2.2}$$

The latter is at most $L(q-1)+\lfloor 2\sqrt{q}\rfloor - 1$ for all $q \geq 19$.

Now suppose $1 \leq k \leq L-2$. First, assume $I(P_i) = 0$ for all $1 \leq i \leq k$. Then by Corollary 2.2 (2),

$$Z(f) \leq s(q+1) + (k-s)(q-1) = 2s + k(q-1),$$

where s is the number of twice unit triangles among the P_i. Since the sum of the full Minkowski lengths of the P_i cannot exceed L we have $2s+(k-s) \leq L$, i.e., $s \leq L-k$. Using this inequality along with $k \leq L-2$, we obtain

$$Z(f) \leq 2s + k(q-1) \leq 2L + k(q-3) \leq (L-2)(q-1) + 4.$$

The latter is at most $L(q-1)$ for all $q \geq 3$ and the bounds follow.

Suppose $I(P_i) > 0$ for at least one of the P_i. Then, as we will show in Lemma 2.7,

$$Z(f) \leq k(q-1) + 2\big(A + 3/2 - 2k\big)\sqrt{q} + 2. \tag{2.3}$$

Now the right-hand side will be at most $L(q-1)+2\sqrt{q}-1$ whenever q satisfies

$$(L-k)q - 2(A+1/2-2k)\sqrt{q} - (L-k+3) \geq 0. \tag{2.4}$$

Before proceeding we introduce the following notation: $m = L-k$, $d = A/2-L+1/4$. Then (2.4) becomes

$$mq - 4(d+m)\sqrt{q} - (m+3) \geq 0, \quad 2 \leq m \leq L-1.$$

Since this is a quadratic inequality in $\sqrt{q}$, it will hold if

$$\sqrt{q} \geq C + \sqrt{C^2 + 1 + 3/m}, \quad \text{where } C = 2 + 2d/m.$$

Since $m \geq 2$, it is enough to choose $\sqrt{q} \geq C + \sqrt{C^2 + 5/2}$. Finally, if $d \geq 0$, then $C \leq 2 + d$, since $m \geq 2$, and it is enough to choose

$$q \geq \left(c + \sqrt{c^2 + 5/2}\right)^2, \quad \text{where} \quad c = 2 + d = A/2 - L + 9/4.$$

If $d < 0$, then $C < 2$ and it is enough to choose $q \geq 23$.

(2) The proof of the second statement is completely analogous. First, if f has L irreducible factors, then the bound holds by Proposition 2.3, (3). Second, if f has fewer than L factors we choose q large enough so that the right-hand sides of (2.2) and (2.3) are no greater than $L(q-1)$. The same arguments as before show that it is enough to choose

$$q \geq \max\left(37, \left(c + \sqrt{c^2 + 2}\right)^2\right), \quad \text{where} \quad c = A/2 - L + 11/4. \qquad \square$$

It remains to prove the following lemma.

LEMMA 2.7. *Let* $f = f_1 \cdots f_k$, *for* $1 \leq k \leq L-2$, *and* $I(P_i) > 0$ *for at least one* i. *Then*

$$Z(f) \leq k(q-1) + 2\big(A + 3/2 - 2k\big)\sqrt{q} + 2.$$

Proof. We order the P_i so that for $1 \leq i \leq t$ every P_i either has interior lattice points or is twice a unit triangle. Then, according to Proposition 2.1 and Corollary 2.2, we have

$$Z(f) \leq t(q+1) + 2\sqrt{q}\sum_{i=1}^{t} I(P_i) + (k-t)(q-1). \tag{2.5}$$

Now we want to get a bound for $\sum_{i=1}^{t} I(P_i)$. Recall that given two polytopes Q_1 and Q_2 in $\mathbb{R}^2$, their normalized *mixed volume* (two-dimensional) is

$$V(Q_1, Q_2) = A(Q_1 + Q_2) - A(Q_1) - A(Q_2).$$

The mixed volume is symmetric; bilinear with respect to Minkowski addition; monotone increasing (i.e., if $Q_1' \subset Q_1$, then $V(Q_1', Q_2) \leq V(Q_1, Q_2)$); and $\mathrm{AGL}(2,\mathbb{Z})$-invariant (see, for example, [4, p. 138]). This implies that

$$V(P_i, P_j) \geq 2 \quad \text{for } 1 \leq i \leq t \quad \text{and } 1 \leq j \leq k. \tag{2.6}$$

Indeed, by monotonicity it is enough to show that $V(P_i, E) \geq 2$ for any lattice segment E, and by $\mathrm{AGL}(2,\mathbb{Z})$-invariance we can assume that E is horizontal. It follows readily from the definition that $V(P_i, E) = h(P_i)|E|$, where $h(P_i)$ is the length of the horizontal projection of P_i (the height of P_i) and $|E|$ is the length of E. Clearly, $|E| \geq 1$ and $h(P_i) \geq 2$ if P_i has at least one interior lattice point or is twice a unit triangle.

Using (2.6) and bilinearity of the mixed volume, by induction we obtain

$$A \geq A\Big(\sum_{i=1}^{k} P_i\Big) = A(P_1) + A\Big(\sum_{i=2}^{k} P_i\Big) + V\Big(P_1, \sum_{i=2}^{k} P_i\Big)$$

$$\geq A(P_1) + A\Big(\sum_{i=2}^{k} P_i\Big) + 2(k-1) \geq \cdots$$

$$\geq \sum_{i=1}^{t} A(P_i) + A\Big(\sum_{i=t+1}^{k} P_i\Big) + 2\sum_{i=1}^{t}(k-i) \geq \sum_{i=1}^{t} A(P_i) + 2kt - t^2 - t.$$

Now, by Pick's formula $A(P_i) = I(P_i) + \frac{1}{2}B(P_i) - 1 \geq I(P_i) + \frac{1}{2}$ since $B(P_i)$, the number of boundary lattice points, is at least 3. Therefore

$$\sum_{i=1}^{t} I(P_i) \leq A + t^2 + \frac{t}{2} - 2kt.$$

Substituting this into (2.5) and simplifying, we obtain

$$Z(f) \leq k(q-1) + 2\sqrt{q}\left(A + t^2 + \frac{t}{2} - 2kt\right) + 2t. \tag{2.7}$$

It remains to note that the maximum of the right-hand side of (2.7) is attained at $t = 1$, provided $k \geq 1$ and $q \geq 4$, which establishes the required inequality. □

3. Two algorithms. Given a polytope P, to make use of our bounds in Theorem 2.6 it remains to understand

(1) how to find $L(P)$, the full Minkowski length of P, and
(2) how to determine whether there is a maximal Minkowski decomposition in P one of whose summands is an exceptional triangle.

Here we provide algorithms that answer these questions in polynomial time in $|P \cap \mathbb{Z}^2|$.

Recall that a zonotope $Z = \sum_{i=1}^{k} E_j \subseteq P$ is called *maximal for* P if k, the number of nontrivial Minkowski summands (counting their multiplicities), is equal to $L(P)$.

It follows from Proposition 1.2 that a maximal zonotope always exists although it is usually not unique. It turns out that any maximal zonotope of P has at most four distinct summands and among them there are maximal zonotopes with a particularly easy description.

PROPOSITION 3.1. *Let P be a lattice polygon. Then we have the following:*

(1) *Any zonotope Z maximal for P has at most 4 different summands.*
(2) *There exists a zonotope Z maximal for P with at most 3 different summands. Moreover, up to an $\mathrm{AGL}(2,\mathbb{Z})$-transformation these summands are $[0, e_1]$, $[0, e_2]$, and $[0, e_1 + e_2]$.*

Proof. Let $Z = \sum_{i=1}^{L} E_j$ be a zonotope maximal for P, and let v_j be the vector of E_j. According to Lemma 1.7, $|\det(v_i, v_j)| \leq 2$ for any $1 \leq i, j \leq k$.

The case when all v_i are the same is trivial. Suppose there are exactly two different summands; i.e., $Z = m_1E_1 + m_2E_2$ for some positive integers $m_1 \geq m_2$ and $E_1 \neq E_2$. If $|\det(v_1, v_2)| = 1$, then we can transform (v_1, v_2) to the standard basis (e_1, e_2) and (2) follows. If $|\det(v_1, v_2)| = 2$, then we can assume that $v_1 = e_1$ and $v_2 = e_1 + 2e_2$.

However, $E_1 + E_2$ contains $2[0, e_2]$, therefore we can pass to $Z' = (m_1 - m_2)[0, e_1] + 2m_2[0, e_2]$. Clearly, $Z' \subseteq Z$ and Z' is maximal.

Now suppose that Z has at least three different summands. First, let us assume that $|\det(v_i, v_j)| = 2$ for all $i \neq j$. As before, without loss of generality, $v_1 = e_1$ and $v_2 = e_1 + 2e_2$. Consider $v_3 = (s, t)$. By looking at the determinants $\det(v_i, v_3)$ for $i = 1, 2$, we have $|t| = 2$ and $|t - 2s| = 2$. This implies that v_3 is not primitive, a contradiction. Therefore, $|\det(v_i, v_j)| = 1$ for some $i \neq j$, and we can assume that $v_1 = e_1$ and $v_2 = e_2$. Again, we let $v_3 = (s, t)$ and look at the determinants $\det(v_i, v_3)$ for $i = 1, 2$. We see that the only vectors v_3 (up to central symmetry) that may appear are $(1, 1)$, $(1, -1)$, $(2, 1)$, $(2, -1)$, $(1, 2)$, $(1, -2)$. No two of the last four vectors can appear together as they generate parallelograms of area at least 3. For the same reason $(1, 1)$ cannot appear with $(2, -1)$ or $(1, -2)$, and $(1, -1)$ cannot appear with $(2, 1)$ or $(1, 2)$. We have three possible combinations:

(a) $v_1 = (1, 0), v_2 = (0, 1), v_3 = (1, 1), v_4 = (1, -1)$;
(b) $v_1 = (1, 0), v_2 = (0, 1), v_3 = (1, 1)$, and $v_4 = (1, 2)$ or $v_4 = (2, 1)$;
(c) $v_1 = (1, 0), v_2 = (0, 1), v_3 = (1, -1)$, and $v_4 = (1, -2)$ or $v_4 = (2, -1)$.

We have proved our first claim. To prove the second, note that we can actually reduce the number of distinct segments E_j. In case (a), $2E_1 \subset E_3 + E_4$, and we will be able to get rid of either E_3 or E_4 by replacing $E_3 + E_4$ with $2E_1$. In either case, the remaining segments are $\mathrm{AGL}(2, \mathbb{Z})$-equivalent to $[0, e_1]$, $[0, e_2]$, and $[0, e_1 + e_2]$.

In case (b) we can assume that $v_4 = (1, 2)$. Since $2E_2 \subset E_1 + E_4$, we will be able to get rid of either E_1 or E_4 and the remaining segments are $\mathrm{AGL}(2, \mathbb{Z})$-equivalent to $[0, e_1]$, $[0, e_2]$, and $[0, e_1 + e_2]$. Case (c) is obtained from (b) by flipping the second coordinate. ∎

To find $L(P)$ we only need to look at all of the zonotopes $Z \subseteq P$ with at most three different summands $\mathrm{AGL}(2, \mathbb{Z})$-equivalent to $[0, e_1]$, $[0, e_2]$, and $[0, e_1 + e_2]$ and find the one that has the largest number of summands (counting multiplicities).

THEOREM 3.2. *Let P be a lattice polygon, and let $|P \cap \mathbb{Z}^2|$ be the number of lattice points in P. Then the full Minkowski length $L(P)$ can be found in polynomial time in $|P \cap \mathbb{Z}^2|$.*

Proof. The case when P is one-dimensional is trivial so we will be assuming that P has dimension two.

For every triple of points $\{A, B, C\} \subseteq P \cap \mathbb{Z}^2$, where it is important which point goes first and the order of the other two does not matter, we check if $E_1 = [A, B]$ and $E_2 = [A, C]$ generate a parallelogram of area one. If so, we want to construct various zonotopes whose summands are E_1, E_2, and $E_3 = [A, B + C]$. We do this in the most straightforward way.

First, for every $1 \leq i \leq 3$, we find M_i, the largest integer such that a lattice translate of $M_i E_i$ is contained in P. For this we find the maximum number of lattice points in the linear sections of P with lines in the direction of E_i (there are finitely many such lines with at least one lattice point of P).

Second, for each triple of integers $m = (m_1, m_2, m_3)$, where $0 \leq m_i \leq M_i$, we check if some lattice translate of the zonotope $Z_m = m_1 E_1 + m_2 E_2 + m_3 E_3$ is contained in P (we run through lattice points D in P to check if $D + Z_m$ is contained in P). For all such zonotopes that fit into P we look at $m_1 + m_2 + m_3$ and find the maximal possible value M of this sum.

Finally, the largest such sum M over all choices of $\{A, B, C\} \subseteq P \cap \mathbb{Z}^2$ is $L(P)$, by Proposition 3.1. Clearly, this algorithm is polynomial in $|P \cap \mathbb{Z}^2|$.

Notice that in the previous argument we have taken care of the maximal zonotopes that are possibly multiples of a single segment. Indeed, if $[A, B]$ is a primitive segment connecting two lattice points in P, then unless P is one-dimensional there is a lattice point C in P such that $[A, B]$ and $[A, C]$ generate a parallelogram of area one. We can assume that A is the origin and $B = (1, 0)$. Let $C = (k, l)$ be a lattice point in P with smallest positive l (flip P with respect to the x-axis if necessary). By the minimality of l the triangle ABC has no lattice points except its vertices. By Pick's formula, its area is $1/2$ and we have found the required third vertex C. ∎

THEOREM 3.3. *Let P be a lattice polygon in $\mathbb{R}^n$. Then we can decide in polynomial time in $|P \cap \mathbb{Z}^2|$ if there is a maximal Minkowski decomposition in P one of whose summands is an exceptional triangle.*

Proof. We first run the algorithm from Theorem 3.2 to find $L(P)$. Next, for each triple of points $A, B, C \in P \cap \mathbb{Z}^2$ we check if the triangle T_{ABC} has exactly four lattice points—the three vertices A, B, C and one point D strictly inside the triangle. If so, this triangle is exceptional. If this triangle is a summand in some maximal Minkowski decomposition in P, then the other summands that may appear in this decomposition are the primitive segments E_1, E_2, and E_3 connecting D to the vertices A, B, C (see Remark 1.9).

Now it remains to look at all Minkowski sums $T_{ABC} + m_1E_1 + m_2E_2 + m_3E_3$ with $m_1 + m_2 + m_3 = L(P) - 1$ and check if any of them fits into P. If this indeed happens for some T_{ABC}, then there is a maximal decomposition in P with an exceptional triangle. Otherwise any maximal decomposition is a sum of primitive segments and unit triangles. Clearly, this algorithm is polynomial in $|P \cap \mathbb{Z}^2|$. ∎

4. Three examples. In this section we illustrate our methods with three examples. Example 2 was given by Joyner in [8]. Example 3 appears in Little and Schenck's paper [10].

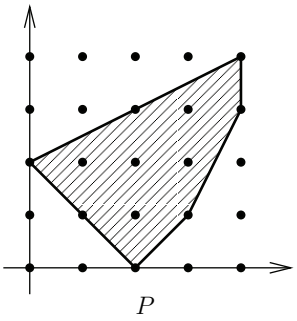


FIG. 7. *Pentagon.*

Example 1. Consider the pentagon P with vertices $(2, 0)$, $(0, 2)$, $(4, 4)$, $(4, 3)$, and $(3, 1)$ as in Figure 7. One can easily check that $L(P) = 3$ and there is a maximal decomposition in P containing T_0 (in fact, P contains $T_0 + [0, e_2] + [0, e_1 + e_2]$). Note that P defines a toric surface code of dimension $n = |P \cap \mathbb{Z}^2| = 12$. To apply Theorem 2.6 we compute $A = 15/2$, so $c = 3$. Therefore,

$$d(C_P) \geq (q-1)^2 - 3(q-1) - 2\sqrt{q} + 1$$

for all $q \geq 41$. In this particular example we can establish a better lower bound for q, namely $q \geq 19$. Indeed, we have already seen in the proof of Theorem 2.6 that every

f with 2 absolutely irreducible factors will have at most $3(q-1)+2\sqrt{q}-1$ zeros for all $q \geq 19$ (see (2.2)). If f is absolutely irreducible, then we use (2.1). Then it has at most $q+1+\lfloor 10\sqrt{q} \rfloor - 2$ zeros since $P_f \subseteq P$ has at most 5 interior lattice points in which case it will have at least two primitive edges. It remains to notice that

$$q+1+\lfloor 10\sqrt{q} \rfloor - 2 \leq 3(q-1)+2\sqrt{q}-1$$

for all $q \geq 19$.

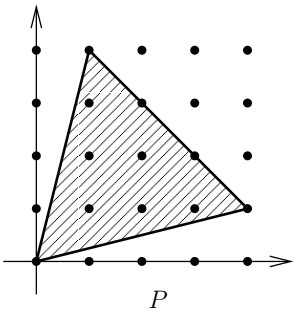


FIG. 8. *Triangle.*

Example 2. Consider the triangle P with vertices $(0,0)$, $(4,1)$, and $(1,4)$ (see Figure 8). This example is similar to the previous one. We also have $L(P) = 3$, $A = 15/2$, but the dimension of the corresponding toric surface code is slightly smaller, $n = |P \cap \mathbb{Z}^2| = 11$. However, in this case P has no exceptional triangles in any maximal decomposition. Therefore, Theorem 2.6 provides a better bound for the minimum distance

$$d(C_P) \geq (q-1)^2 - 3(q-1),$$

which holds for all $q \geq 53$. As before, this can be improved to $q \geq 37$ using (2.1) and the fact that $I(P) = 6$. Note that $f = xy(x-a)(x-b)(x-c)$, for $a, b, c \in \mathbb{F}_q^*$ distinct, has exactly $3(q-1)$ zeros in $(\mathbb{F}_q^*)^2$, hence for $q \geq 37$ the above bound is exact

$$d(C_P) = (q-1)^2 - 3(q-1). \tag{4.1}$$

For $q = 8$ this was previously established by Joyner [8]. Also (4.1) follows from Little and Schenck's result [10] for all $q \geq (4I(P)+3)^2 = 729$.

Example 3. Let P be the hexagon with vertices $(1,0)$, $(0,1)$, $(1,2)$, $(3,3)$, $(3,2)$, and $(2,0)$ (see Figure 9). We have $L(P) = 3$, $A = 5$, and $\mathcal{C}_P$ has dimension nine. Also P has no maximal decomposition with an exceptional triangle. Therefore, Theorem 2.6 implies

$$d(C_P) \geq (q-1)^2 - 3(q-1)$$

for all $q \geq 37$. Little and Schenck's result [10] proves this bound for $q > 225$. In fact we can show more in this example: for all $q \geq 11$

$$d(C_P) = (q-1)^2 - 3(q-1) + 2. \tag{4.2}$$

To see this, first note that $f = x(x-a)(y-b)(y-c)$, for $a, b, c \in \mathbb{F}_q^*$ distinct, has exactly $3(q-1)-2$ zeros in $(\mathbb{F}_q^*)^2$. Furthermore, every maximal decomposition in P

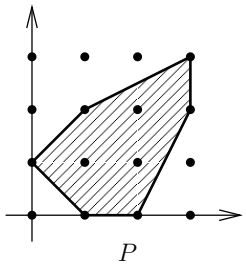


FIG. 9. *Hexagon.*

is of the form $E_1 + 2E_2$, where E_i is a primitive segment in the direction of e_1, e_2, or $e_1 + e_2$. This implies that every polynomial f with the largest number of absolutely irreducible factors (three) will have at most $3(q-1)-2$ zeros in $(\mathbb{F}_q^*)^2$ (here we take into account the intersections of the irreducible curves defined by the factors of f).

Now we claim that for $q \geq 11$ polynomials with fewer factors (one or two) will have at most $3(q-1)-2$ zeros in $(\mathbb{F}_q^*)^2$ as well. Indeed, decompositions with two summands in P can have at most one exceptional triangle, hence, $Z(f) \leq 2(q-1) + \lfloor 2\sqrt{q} \rfloor$ for every f with two irreducible factors. This will be no greater than $3(q-1)-2$ for $q \geq 9$. If f is absolutely irreducible, then by (2.1) $Z(f) \leq q + 1 + \lfloor 6\sqrt{q} \rfloor - 3$, which is no greater than $3(q-1)-2$ starting with $q = 11$.

The computations preformed in [10] show the validity of (4.2) for all $5 \leq q \leq 11$ except for $q = 8$ when the answer is $d(C_P) = (q-1)^2 - 3(q-1) = 28$. For example, the polynomial $x^2 + y + x^3y^3$ has exactly 21 zeros in $(\mathbb{F}_8^*)^2$, and so the corresponding codeword has weight 28. We now have a complete understanding of this example.

Acknowledgments. We thank Leah Gold and Felipe Martins for helpful discussions on coding theory.

REFERENCES

[1] Y. AUBRY AND M. PERRET, *A Weil theorem for singular curves*, in Arithmetic, Geometry and Coding Theory (Luminy, 1993), de Gruyter, Berlin, 1996, pp. 1–7.

[2] V. BATYREV AND B. NILL, *Multiples of lattice polytopes without interior lattice points*, Mosc. Math. J., 7 (2007), pp. 195–207, 349.

[3] M. BECK AND S. ROBINS, *Computing the continuous discretely. Integer-point enumeration in polyhedra*, Undergraduate Texts in Mathematics, Springer, New York, 2007.

[4] YU. D. BURAGO AND V. A. ZALGALLER, *Geometric Inequalities*, Springer-Verlag, Berlin, 1988.

[5] W. FULTON, *Introduction to Toric Varieties*, Ann. of Math. Stud. 131, Princeton University Press, Princeton, NJ, 1993.

[6] J. HANSEN, *Toric surfaces and error-correcting codes*, in Coding Theory, Cryptography, and Related Areas, Springer, Berlin, 2000, pp. 132–142.

[7] A. G. HOVANSKII, *Newton polyhedra, and the genus of complete intersections*, Funktsional. Anal. i Prilozhen., 12 (1978), pp. 51–61 (in Russian).

[8] D. JOYNER, *Toric codes over finite fields*, Appl. Algebra Engrg. Comm. Comput., 15 (2004), pp. 63–79.

[9] A. G. KHOVANSKII, *Newton polytopes, curves on toric surfaces, and inversion of Weil's theorem*, Russian Math. Surveys, 52 (1997), pp. 1251–1279.

[10] J. LITTLE AND H. SCHENCK, *Toric surface codes and Minkowski sums*, SIAM J. Discrete Math., 20 (2006), pp. 999–1014.

[11] M. TSFASMAN, S. VLĂDUŢ, AND D. NOGIN, *Algebraic Geometric Codes: Basic Notions*, Mathematical Surveys and Monographs 139, AMS, Providence, RI, 2007.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 401–406

REAL ZEROS AND NORMAL DISTRIBUTION FOR STATISTICS ON STIRLING PERMUTATIONS DEFINED BY GESSEL AND STANLEY*

MIKLÓS BÓNA†

Abstract. We study Stirling permutations defined by Gessel and Stanley in [*J. Combin. Theory Ser. A*, 24 (1978), pp. 25–33]. We prove that their generating function according to the number of descents has real roots only. We use that fact to prove that the distribution of these descents and other equidistributed statistics on these objects converge to a normal distribution.

Key words. permutations, multisets, descents, normal distribution, real zeros

AMS subject classifications. 05A05, 05A15, 05A16

DOI. 10.1137/070702254

1. Introduction. In [8], Ira Gessel and Richard Stanley defined an interesting class of multiset permutations called *Stirling permutations*. Let Q_n denote the set of all permutations of the multiset $\{1,1,2,2,\ldots,n,n\}$ in which, for all i, all entries between the two occurrences of i are larger than i. For instance, Q_2 has three elements, namely, 1122, 1221, and 2211. It is not difficult to see that Q_n has $1\cdot 3\cdot\cdots\cdot(2n-1) = (2n-1)!!$ elements. Gessel and Stanley then proved many enumerative results for these permutations and showed several connections between these and other combinatorial objects, such as set partitions.

Counting Stirling permutations by descents, the authors of [8] found a recurrence relation similar to the recurrence relation known for classic permutations. In this paper, we will continue in that direction. First, we show the simple but interesting fact that on Q_n the descent and the *plateau* statistics, to be defined in the next section, are equidistributed. Then we prove that, for any fixed n, the generating polynomial of all Stirling permutations in Q_n with respect to the descent statistic has real roots only. This is analogous to the well-known case (see Theorem 1.33 of [1]) of classic permutations, namely, the result that all the roots of Eulerian polynomials are real. Finally, we apply a classic result of Bender to use this real roots property to prove that the descents of Stirling permutations in Q_n are normally distributed.

2. Stirling permutations and real zeros. Let $q = a_1a_2\cdots a_{2n} \in Q_n$ be a Stirling permutation. Let the index i be called an *ascent* of q if $i = 0$ or $a_i < a_{i+1}$, let i be called a *descent* of q if $i = 2n$ or $a_i > a_{i+1}$, and let i be called a *plateau* of q if $a_i = a_{i+1}$. It is obvious that the ascent and descent statistics are equidistributed, since reversing an element of Q_n turns ascents into descents and vice versa. It is somewhat less obvious that the plateau statistic is also equidistributed with the previous two. This fact, and a reason for it, are the content of the next proposition. Note that its first identity, (1), was proved in [8].

PROPOSITION 1. *Let $C_{n,i}$ be the number of elements of Q_n with i descents. Then for all positive integers $n, i \geq 2$, we have*

$$C_{n,i} = iC_{n-1,i} + (2n-i)C_{n-1,i-1}. \tag{1}$$

*Received by the editors September 7, 2007; accepted for publication (in revised form) September 13, 2008; published electronically January 14, 2009.
http://www.siam.org/journals/sidma/23-1/70225.html

†Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 (bona@math.ufl.edu). This author was partially supported by an NSA Young Investigator Award.

Similarly, let $c_{n,i}$ be the number of elements of Q_n with i plateaux. Then for all positive integers $n, i \geq 2$, we have

$$c_{n,i} = ic_{n-1,i} + (2n-i)c_{n-1,i-1}. \tag{2}$$

In particular, since $C_{1,1} = c_{1,1} = 1$ and $C_{1,0} = c_{1,0} = 0$, the identity

$$C_{n,i} = c_{n,i} \tag{3}$$

holds.

Proof. There are two ways to obtain an element of Q_n from an element $p \in Q_{n-1}$ by inserting two copies of n into consecutive positions. Either p must have i descents, and then we insert the two copies of n into a descent, or p has $i-1$ descents, and then we insert the two consecutive copies of n into one of the $(2n-1)-(i-1) = 2n-i$ positions that are not descents.

The argument proving (2) is analogous. □

COROLLARY 1. *On average, elements of Q_n have $(2n+1)/3$ ascents, $(2n+1)/3$ descents, and $(2n+1)/3$ plateaux.*

Proposition 1 enables us to prove a strong result on the roots of the polynomials $\sum_{i=1}^{n} C_{n,i}x^i$. The method we use follows an idea of Wilf [9], [1, Theorem 1.33], who used it on classic permutations.

THEOREM 1. *Let $C_n(x) = \sum_{i=1}^{n} C_{n,i}x^i$. Then for all positive integers n, the roots of the polynomial $C_n(x)$ are all real, distinct, and nonpositive.*

Proof. For $n = 1$, one sees that $C_1(x) = x$, and the statement holds. For $n = 2$, one sees that $C_2(x) = 2x^2 + x = x(2x+1)$, and so the statement again holds.

For $n \geq 3$, recurrence relation (1) implies

$$C_n(x) = (x - x^2)C'_{n-1}(x) + (2n-1)xC_{n-1}(x), \tag{4}$$

as can be seen by equating coefficients of x^i. The right-hand side is similar to the derivative of a product, which suggests the following rearrangement:

$$C_n(x) = x(1-x)^{2n}\frac{d}{dx}\left((1-x)^{1-2n}C_{n-1}(x)\right). \tag{5}$$

Let us now assume inductively that the roots of $C_{n-1}(x)$ are real, distinct, and nonpositive. Clearly, $C_n(x)$ vanishes at $x = 0$. Furthermore, by Rolle's theorem, (5) shows that $C_n(x)$ has a root between any pair of consecutive roots of $C_{n-1}(x)$. This counts for $n-1$ roots of $C_n(x)$. So the last root must also be real, since complex roots of polynomials with real coefficients must come in conjugate pairs.

There remains to show that the last root of $C_n(x)$ must be on the right of the rightmost root of C_{n-1}. Consider (4) at the rightmost root x_0 of C_{n-1}. As x_0 is negative, we know that $x_0 - x_0^2 < 0$, and so $C_n(x_0)$ and $C'_{n-1}(x_0)$ have opposite signs. The claim now follows, since in $-\infty$, the polynomials $C_n(x)$ and $C'_{n-1}(x)$ must converge to the same (infinite) limit as their degrees are of the same parity. As $C'_{n-1}(x)$ has no more roots on the right of x_0, the polynomial $C_n(x)$ must have one. □

Note that we have in fact proved that the roots of $C_{n-1}(x)$ and $C_n(x)$ are interlacing, so the sequence $C_1, C_2, \ldots$ is a *Sturm sequence.*

As an immediate application of the real zeros property, we can determine where peak (or peaks) of the sequence $C_{n,1}, C_{n,2}, \ldots, C_{n,n}$ is. Our tool in doing so is the following theorem of Darroch.

THEOREM 2 (Darroch [4]). *Let $A(x) = \sum_{k=0}^{n} a_k x^k$ be a polynomial that has real roots only that satisfies $A(1) > 0$. Let m be an index so that $a_m = \max_{0 \le i \le n} a_i$. Let $\mu = A'(1)/A(1)$. Then we have*

$$|\mu - m| < 1.$$

In particular, a sequence with the real zeros property can have at most two peaks. Note that $A'(1) = \sum_{i=0}^{n} i a_i$ and $A(1) = \sum_{i=0}^{n} a_i$; therefore, $A'(1)/A(1)$ is nothing else but the weighted average of the coefficients a_i, with i being the weight of a_i. So in the particular case when $A(x) = C_n(x)$, we have

$$\begin{aligned} \frac{C_n'(1)}{C_n(1)} &= \frac{\sum_i i C_{n,i}}{\sum_i C_{n,i}} \\ &= \sum_i i \cdot \frac{C_{n,i}}{(2n-1)!!} \\ &= \frac{2n+1}{3}, \end{aligned}$$

where the last step follows from Corollary 1. Indeed, $\frac{C_{n,i}}{(2n-1)!!}$ is just the probability that a randomly selected Stirling permutation of length n has exactly i descents, so $sum_i i \cdot \frac{C_{n,i}}{(2n-1)!!}$ is just the expected number of descents in such permutations.

Therefore, by Theorem 2, we obtain the following result.

THEOREM 3. *Let i be an index so that $C_{n,i} = \max_k C_{n,k}$. Then*

1. *$i = (2n+1)/3$ if $(2n+1)/3$ is an integer, and*
2. *$i = \lfloor (2n+1)/3 \rfloor$ or $i = \lceil (2n+1)/3 \rceil$ if $(2n+1)/3$ is not an integer.*

3. Stirling permutations and normal distribution. In this section, we prove that the plateaux (equivalently, ascents; equivalently, descents) of Stirling permutations are normally distributed. Our main tool is the following result of Bender. Let X_n be a random variable, and let $a_n(k)$ be a triangular array of nonnegative real numbers, $n = 1, 2, \ldots$, and $1 \le k \le m(n)$ so that

$$P(X_n = k) = p_n(k) = \frac{a_n(k)}{\sum_{i=1}^{m(n)} a_n(i)}.$$

Set $g_n(x) = \sum_{k=1}^{m(n)} p_n(k) x^k$.

We need to introduce some notation for transforms of the random variable Z. Let $\bar{Z} = Z - E(Z)$, let $\tilde{Z} = \bar{Z}/\sqrt{\mathrm{Var}(Z)}$, and let $Z_n \to N(0,1)$ mean that Z_n converges in distribution to the standard normal variable.

THEOREM 4 (see [2]). *Let X_n and $g_n(x)$ be as above. If $g_n(x)$ has real roots only, and*

$$\sigma_n = \sqrt{Var(X_n)} \to \infty,$$

then $\tilde{X}_n \to N(0,1)$.

See [3] for related results.

We want to use Theorem 4 to prove that the plateaux of permutations in Q_n are normally distributed. Because of Theorem 1, all we need for that is to prove that the variance of the number of these plateaux converges to infinity as n goes to infinity. We will accomplish more by proving an explicit formula for this variance. In order to state that formula, let $Y_{n,i}$ be the indicator random variable of the event that in

a randomly selected element of Q_n, the two copies of i are consecutive; that is, they form a plateau. Note that $P(Y_{n,n} = 1) = E(Y_{n,n}) = 1$. Set $Y_n = \sum_{i=1}^{n} Y_{n,i}$.

THEOREM 5. *For all positive integers n, the equality*

$$Var(Y_n) = \frac{2n^2 - 2}{18n - 9} \tag{6}$$

holds.

Proof. We are going to use the identity $\mathrm{Var}(Y_n) = E(Y_n^2) - E(Y_n)^2$. We have seen in Corollary 1 that $E(Y_n) = \frac{2n+1}{3}$. Let $s_n = E(Y_n^2)$. The key element of our computations is the following lemma.

LEMMA 1. *For all positive integers n, the equality*

$$s_{n+1} = \frac{2n-1}{2n+1} \cdot s_n + \frac{4n+4}{3} \tag{7}$$

holds.

Proof. In order to prove (7), we need the following simple facts.

PROPOSITION 2.

1. *For all positive integers n, and all indices $i \neq j$ that satisfy $1 \leq i, j \leq n$, the equality*
$$E(Y_{n+1,i}Y_{n+1,j}) = \frac{2n-1}{2n+1} E(Y_{n,i}Y_{n,j})$$
holds.
2. *For all positive integers n and all indices $1 \leq i \leq n$, the equality*
$$E(Y_{n+1,i}) = \frac{2n}{2n+1} E(Y_{n,i})$$
holds.
3. *For all indices $i \leq n+1$, the equality*
$$E(Y_{n+1,i}Y_{n+1,n+1}) = E(Y_{n+1,i})$$
holds. In particular, $E(Y_{n+1,n+1}) = 1$.

Proof.

1. In order to get an element of Q_{n+1} in which i and j are both plateaux, take an element of Q_n in which i and j are both plateaux, and insert two consecutive copies of $n+1$ into any of the $2n-1$ available places, that is, anywhere but between the two copies of i or the two copies of j.
2. In order to get an element of Q_{n+1} in which i is a plateau, insert two consecutive copies of $n+1$ into any of the $2n$ available slots, that is, anywhere but between the two copies of i.
3. This part of the proof is obvious since $n+1$ is always a plateau in elements of Q_{n+1}. □

We return to proving Lemma 1.

Note that $s_{n+1} = \sum_{1 \leq i,j \leq n+1} E(Y_{n+1,i}Y_{n+1,j})$. The latter can be split into partial sums, based on whether i or j are equal to $n+1$, as follows:

$$s_{n+1} = \sum_{1 \leq j \leq n+1} E(Y_{n+1,n+1}Y_{n+1,j}) + \sum_{1 \leq i \leq n} E(Y_{n+1,i}Y_{n+1,n+1})$$

$$+ \sum_{1 \leq i,j \leq n} E(Y_{n+1,i}Y_{n+1,j}).$$

Based on part 3 of Proposition 2, this simplifies to

$$s_{n+1} = \sum_{1\le j\le n+1} E(Y_{n+1,j}) + \sum_{1\le i\le n} E(Y_{n+1,i}) + \sum_{\substack{1\le i,j\le n\\ i\ne j}} E(Y_{n+1,i}Y_{n+1,j})$$

$$+ \sum_{1\le i\le n} E(Y_{n+1,i}).$$

Now note that the first sum on the right-hand side is just $E(Y_{n+1})$, and the second sum is $E(Y_{n+1} - Y_{n+1,n+1}) = E(Y_{n+1}) - 1$; use part 1 of Proposition 2 on the third sum and part 2 of Proposition 2 on the fourth sum to get

$$s_{n+1} = 2E(Y_n) - 1 + \frac{2n-1}{2n+1}\left(s_n - E(Y_n)\right) + \frac{2n}{2n+1}E(Y_n).$$

Recalling from Corollary 1 that $E(Y_n) = \frac{2n+1}{3}$, this reduces to (7). □

Using the recursive formula proved in Lemma 1, it is routine to prove that

$$s_n = E(Y_n^2) = \frac{8n^3 + 6n^2 - 2n - 3}{18n - 9}. \tag{8}$$

Therefore, $\mathrm{Var}(Y_n) = s_n - E(Y_n)^2 = \frac{2n^2-2}{18n-9}$, as claimed. □

THEOREM 6. *The distribution of the number of plateaux of elements of Q_n converges to a normal distribution as n goes to infinity. That is, $\tilde{Y}_n \to N(0,1)$.*

Proof. Let $X_n = Y_n$, and let $g_n(x) = \frac{1}{(2n-1)!!}C_n(x)$. Then Theorems 1 and 5 show that the conditions of Theorem 4 are satisfied, and the claim follows from Theorem 4. □

4. Remarks. Corollary 1 shows that $E(Y_n) = (2n+1)/3$. It is not difficult to prove that $E(Y_{n,n-i}) = \prod_{j=1}^{i} \frac{2n-2j}{2n-2j+1}$. By the linearity of expectation this proves the interesting identity

$$\sum_{i=0}^{n-1}\prod_{j=1}^{i}\frac{2n-2j}{2n-2j+1} = \frac{2n+1}{3},$$

where the empty product (indexed by $i = 0$) is considered to be 1.

The proof of the equidistribution of the descent and plateau statistics we gave is very simple, but it is of recursive nature. It can be used to define an algorithm that recursively constructs a bijection f from the set of permutations in Q_n that have k descents into the set of permutations in Q_n that have k plateaux. Let us assume that such a bijection has already been constructed for Q_{n-1} and any $k \le n-1$. If $p \in Q_n$, and p has k descents, then let $p' \in Q_{n-1}$ be the permutation obtained from p by removing the two copies of n. Let q' be the image of p' under the bijection already constructed for Q_{n-1}. If p is obtained from p' by inserting two copies of n into the ith descent of p', then let $f(p) = q$ be the permutation obtained from q' by inserting the two copies of n into the ith plateau of q. If p is obtained from p' by inserting two copies of n into the jth nondescent of p', then let $f(p) = q$ be the permutation obtained from q' by inserting the two copies of n into the jth nondescent of q'.

A direct bijective proof has recently been given by Ju [7].

We mention that the results of this work have recently been extended by Janson [5] and Janson, Kuba, and Panholzer [6].

Acknowledgments. I am indebted to Svante Janson, who pointed out an error in an earlier version of this paper, which led to an improvement of my results. I am grateful to Ira Gessel for having taken the time to show me some earlier unpublished work on the subject.

REFERENCES

[1] M. Bóna, *Combinatorics of Permutations*, Chapman & Hall/CRC, Boca Raton, FL, 2004.
[2] E. A. Bender, *Central and local limit theorems applied to asymptotic enumeration*, J. Combin. Theory Ser. A, 15 (1973), pp. 91–111.
[3] E. R. Canfield, *Central and local limit theorems for coefficients of polynomials of binomial type*, J. Combin. Theory Ser. A, 23 (1977), pp. 275–290.
[4] J. N. Darroch, *On the distribution number of successes in independent trials*, Ann. Math. Statist., 35 (1964), pp. 1317–1321.
[5] S. Janson, *Plane Recursive Trees, Stirling Permutations, and an Urn Model*, preprint; available online at http://arxiv.org/pdf/0803.1129v1.
[6] S. Janson, M. Kuba, and A. Panholzer, *Generalized Stirling Permutations, Families of Increasing Trees, and Urn Models*, preprint; available online at http://arxiv.org/pdf/0805.4804.
[7] H. Ju, *Personal communication*, 2007.
[8] I. Gessel and R. P. Stanley, *Stirling polynomials*, J. Combin. Theory Ser. A, 24 (1978), pp. 25–33.
[9] H. S. Wilf, *Real Zeros of Polynomials That Count Runs and Descending Runs*, Unpublished manuscript, 1998.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 407–427

A BOUND ON THE PATHWIDTH OF SPARSE GRAPHS WITH APPLICATIONS TO EXACT ALGORITHMS*

JOACHIM KNEIS†, DANIEL MÖLLE†, STEFAN RICHTER†, AND PETER ROSSMANITH†

Abstract. We present a bound of $m/5.769 + O(\log n)$ on the pathwidth of graphs with m edges. Respective path decompositions can be computed in polynomial time. Using a well-known framework for algorithms that rely on tree decompositions, this directly leads to runtime bounds of $O^*(2^{m/5.769})$ for MAX-2SAT and MAX-CUT. Both algorithms require exponential space due to dynamic programming. If we agree to accept a slightly larger bound of $m/5.217 + 3$, we even obtain path decompositions with a rather simple structure: all bags share a large set of common nodes. Using branching based algorithms, this allows us to solve the same problems in polynomial space and time $O^*(2^{m/5.217})$.

Key words. graph algorithms, graph theory, algorithms

AMS subject classifications. 05C85, 68R10, 68W01

DOI. 10.1137/080715482

1. Introduction.

In 2005, Fomin and Høie [5, 6] bounded the pathwidth of cubic graphs with n nodes by $(1 + \varepsilon)n/6 + O(\log n)$. Using this result we derive an upper bound of $m/5.769 + O(\log n)$ on the pathwidth of arbitrary graphs.

Combined with a general result on treewidth-based algorithms by Telle and Proskurowski [20], this bound—besides being an interesting graph-theoretical result by itself—implies runtime bounds of $O^*(2^{m/5.769})$ for MAX-2SAT and MAX-CUT, where m is the number of clauses or edges, respectively. The respective algorithms require exponential space because they are based on dynamic programming on tree decompositions. Moreover, we can construct simpler decompositions of width at most $m/5.217 + 3$ that allow for branching-based $O^*(2^{m/5.217})$ algorithms with only polynomial space complexity.

The above runtime bounds are particularly interesting, because all previous results, such as the runtime bounds $O^*(2^{m/2.88})$ [14], $O^*(2^{m/3.44})$ [1], $O^*(2^{m/4})$ [4], and $O^*(2^{m/5})$ [7] for MAX-CUT, lead to much more involved algorithms directly tailored to the problem at hand. In fact, all previous algorithms for MAX-2SAT, MAX-2CSP, and MAX-CUT are based on clever branching and a lot of reduction rules similar to the Davis–Putnam procedure. The $O^*(2^{m/5})$ algorithm for MAX-2SAT [7], for example, employs six reduction rules as well as a six-fold case distinction.

By contrast, the algorithms described in this paper operate on a graph representation of the respective instance (such as the connectivity graph of a 2SAT formula). Nodes of low degree are removed according to a simple set of rules. Otherwise, the algorithms branch on nodes of maximum degree until the graph becomes trivial—namely, three-regular for the bound $O^*(2^{m/5.769})$ or series-parallel for the bound $O^*(2^{m/5.217})$. In general, it does not make a difference which node of maximum degree is selected; each of the two algorithms, however, handles a certain regular case

*Received by the editors February 12, 2008; accepted for publication (in revised form) September 23, 2008; published electronically January 14, 2009. A preliminary version of this paper was presented at the Workshop on Graph Theoretic Concepts in Computer Science (WG 2005) [10].
http://www.siam.org/journals/sidma/23-1/71548.html
†Computer Science Department, RWTH Aachen University, 52056 Aachen, Germany (kneis@cs.rwth-aachen.de, moelle@cs.rwth-aachen.de, richter@cs.rwth-aachen.de, rossmani@cs.rwth-aachen.de). The first author was supported by the DFG under grant RO 927/7.

differently. When proving the first bound, it is vital to avoid cases where a degree-five node has only degree-five neighbors whenever possible. For the second bound, the same holds for degree four.

The reduction rules are very simple and operate on nodes of degree at most two. Nodes of higher degree are removed by branching. A problem can be solved by our framework if all these graph operations correspond to appropriate operations in the problem instance. On the one hand, there need to be corresponding—and efficient—reduction rules for the affected entities represented by single nodes (such as variables in a 2SAT formula that occur only with at most two other variables). On the other hand, it is usually hard to find reduction rules for entities that are represented by a node of high degree; we overcome this problem by simply branching on all possibilities for the entity in question (such as setting a variable to *true* or *false*). The latter operations constitute the expensive part and lead to the exponential runtime bounds.

Our construction results in very intuitive branching algorithms that outperform all previous methods and are also easy to verify and implement. The choice of branching on all possibilities for a basic entity (such as placing a node on the left or right side of a cut) can be seen to bring about the most simple branching possible. The few reduction rules involved are straightforward and can be performed efficiently; the constants and polynomial factors in the resulting runtime bounds are small.

Lately, several authors have matched or improved the aforementioned bounds. Using an argument based on linear programming, Scott and Sorkin presented an alternative proof for the bound of $m/5.769 + O(\log n)$ on the treewidth in 2007 (in fact, their type-III reduction selects nodes in roughly the same way as our algorithm from 2005) [18]. Again, this result leads to runtime bounds of $O^*(2^{m/5.769})$ for Max-2SAT and Max-Cut. Using a similar approach, they also obtained an $O^*(2^{m/5.263})$ algorithm for Max-2SAT and Max-Cut using only polynomial space [17], improving on our bound of $O^*(2^{m/5.217})$. Note that a technical report published shortly before our result already contains this $m/5.263$ bound [16].[1]

Kojevnikov and Kulikov have taken the runtime bound for Max-2SAT under polynomial space restrictions to $O^*(2^{m/5.5})$ [11], and Kulikov and Kutzkov took it subsequently to $O^*(2^{m/5.88})$ [12]. Their algorithm uses a structure similar to our direct algorithm: first the formula is reduced by simple reduction rules, and then they branch on a variable. In order to achieve the improved runtime bounds, Kojevnikov and Kulikov simulate branching on every variable and select the best one for the real branching process, whereas we always select some node of maximum degree. The additional quadratic factor in the runtime vanishes in the O^*-notation but cannot be neglected for practical instances. The current fastest algorithm is due to Raible and Fernau [15] with a runtime of $O^*(2^{m/6.21})$.

Recently, Williams developed an algorithm for Max-2SAT with a runtime bound of only $O^*(2^{2.376n/3})$, depending on fast matrix multiplication. This currently is the fastest algorithm analyzed in the number n of variables [21]. As for approximation results, we refer the reader to [8, 13].

The structure of this paper is as follows. Some preliminaries—particularly regarding treewidth—are detailed in section 2. Section 3 introduces graph reduction rules required for later proofs and establishes an important property, namely, their confluence: no matter in which order the rules are applied, they always lead to the same graph eventually. In section 4, we prove the aforementioned bound of $m/5.769 + O(\log n)$ on

[1] As pointed out by Scott and Sorkin, these bounds can easily be expanded into bounds of $O^*(r^{m/5.263})$ and $O^*(r^{m/5.217})$ for Max-2CSP with r-ary variables.

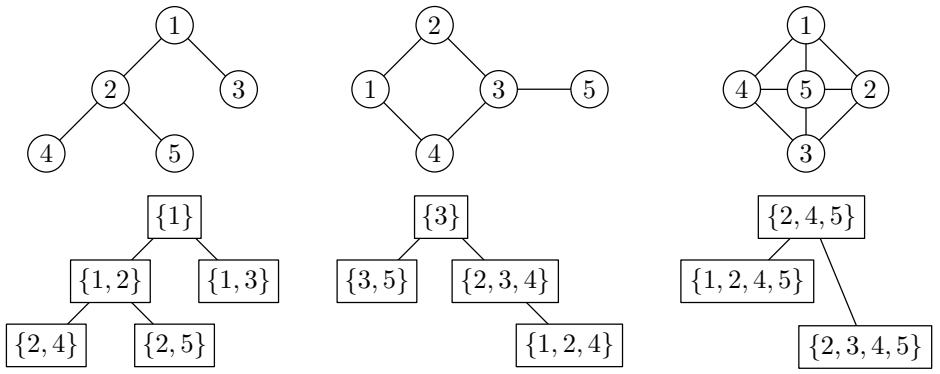


FIG. 2.1. *The three graphs depicted in the upper row have treewidth one, two, and three, respectively. A respective minimal tree decomposition is depicted below each graph.*

the pathwidth of graphs with n nodes and m edges. The construction of polynomial space algorithms for MAX-2SAT and MAX-CUT is discussed in section 5.

2. Preliminaries. Since the notions of treewidth and pathwidth are going to play crucial roles in what follows, we proceed with a brief review of these graph-theoretical concepts. For further reference we recommend the surveys by Bodlaender [2] and Kloks [9].

DEFINITION 2.1. *Let $G = (V, E)$ be a simple, undirected graph. A* tree decomposition *of G is a tree $T = (\mathcal{B}, E_{\mathcal{B}})$, where $\mathcal{B}$ is a family of subsets $B_i \subseteq V$ called* bags *such that*

(i) *each node $v \in V$ occurs in at least one bag $B_i \in \mathcal{B}$,*

(ii) *the two endpoints of each edge $\{v, w\} \in E$ occur in at least one bag $B_i \in \mathcal{B}$ simultaneously, and*

(iii) *for each node $v \in V$, the bags $B_i \in \mathcal{B}$ with $v \in B_i$ induce a connected subgraph of T (i.e., a tree).*

If T is a path, then it is also called a path decomposition *of G.*

Clearly, every graph has a tree decomposition $(\mathcal{B}, T)$ in which $\mathcal{B} = \{V\}$ and T is the degenerated tree consisting of only a single node. In most cases, however, we can find a tree decomposition with smaller bags—only tree decompositions of cliques necessarily have bags that contain all nodes.

DEFINITION 2.2. *The* width *of a tree decomposition is the maximum cardinality of its bags minus one. The* treewidth *of a graph is the minimum width of all tree decompositions of that graph. The* pathwidth *is the minimum width of all path decompositions.*

In particular, trees have treewidth one. See Figure 2.1 for an illustration of treewidth and tree decompositions.

A helpful tool for the analysis of tree decompositions is the so-called robber-and-cops game [19]. If we interpret each bag as a set of positions to be guarded by policemen, then a tree decomposition reveals a strategy to catch a robber who moves along edges at arbitrary speed. The police move according to a traversal of the tree decomposition, but while moving from one node to another, a policeman does not affect the robber in any way. Let us exemplify this game by the second graph from Figure 2.1 and the respective decomposition.

In the beginning, one cop blocks the node labeled 3; the robber can reside only in

the right subgraph (node 5) or in the left subgraph (nodes 1, 2, and 4). The former case allows us to catch the robber by placing a second cop on the node labeled 5. In the latter case, we block the nodes labeled 2 and 4 by placing a second and third cop. This forces the robber to retreat to node 1 such that moving the first cop from node 3 to node 1 ends the chase.

It is easy to see that the pathwidth of a graph is an upper bound for its treewidth. In terms of the robber-and-cops game, a path decomposition of width k corresponds to a strategy for $k+1$ policemen who need to catch an invisible robber.

For the sake of brevity, we define the following symbols and abbreviations for graphs $G = (V, E)$. If $V' \subseteq V$, then $G[V']$ denotes the subgraph of G induced by V'. In a slight abuse of notation, we abbreviate $G[V \setminus V']$ as $G \setminus V'$. Even though we discuss results on simple and undirected graphs, the results also apply to multigraphs with and without loops because treewidth is not affected by multiple edges or loops. As usual, we let n and m denote the number of nodes and edges in a graph, respectively.

3. A confluent set of reduction rules. The upcoming section introduces the graph reduction rules that are to play a crucial role throughout the entire document. Whereas the rules are very simple, we need a few technical arguments to show that the order in which reductions are performed does not affect the outcome. The confluence of the reduction rules is not only an interesting result on its own; it is also a necessary property for the main theorem of this paper: First we show that a specific reduction sequence leads to a tree decomposition of width $m/5.769 + O(\log n)$. Next, we prove that applying the reduction rules in a different order even yields a path decomposition. By the confluence of the reduction rules, the path decomposition is of width $m/5.769 + O(\log n)$ as well.

DEFINITION 3.1. *Let $G = (V, E)$ be a graph and $D \subseteq V$ an arbitrary subset of its nodes. We define the following reduction rules:*

R_0: If there is a $v \notin D$ with $\deg(v) = 0$, then remove v.

R_1: If there is a $v \notin D$ with $\deg(v) = 1$, then remove v.

R_2: If there is a $v \notin D$ with $\deg(v) = 2$, then contract v, i.e., remove v and insert a new edge between its two neighbors, if no such edge exists.

R_D: If G contains a node $v \in D$, then remove v.

R: If any of the above rules can be applied, do so.

R^*: Iterate R as long as possible.

DEFINITION 3.2. *Let $G = (V, E)$ be a graph, let $D \subseteq V$, and let $v \in V$ be a node that can be reduced according to R_0, R_1, R_2, or R_D. Then v is called* reducible *and $G\langle v \rangle$ denotes the graph obtained from G by applying the respective rule on v. For $r \geq 2$ we define $G\langle v_1, \ldots, v_r \rangle = G\langle v_1 \rangle \langle v_2, \ldots, v_r \rangle$ inductively.*

If v_i is reducible in $G\langle v_1, \ldots, v_{i-1} \rangle$ for all $1 \leq i \leq r$, then $(v_1, \ldots, v_r)$ is a valid reduction sequence for G with respect to D. *By ε we denote the (valid) empty reduction sequence. If $G = R^*(G)$, we call G* reduced.

Later, we will need to prove the confluence of these rules. See Figure 3.1 for an example.

LEMMA 3.3. *Let $G = (V, E)$ be a graph, let $D \subseteq V$, and let $x, y \in V$ be two distinct reducible nodes. Then (x, y) and (y, x) are valid reduction sequences for G and $G\langle x, y \rangle = G\langle y, x \rangle$.*

Proof. Deleting or contracting x does not affect the degree of y and vice versa if they are not adjacent. Then (x, y) and (y, x) are both valid reduction sequences and it is easy to see that $G\langle x, y \rangle = G\langle y, x \rangle$.

If x and y are adjacent, then the application of a reduction rule to x or y either

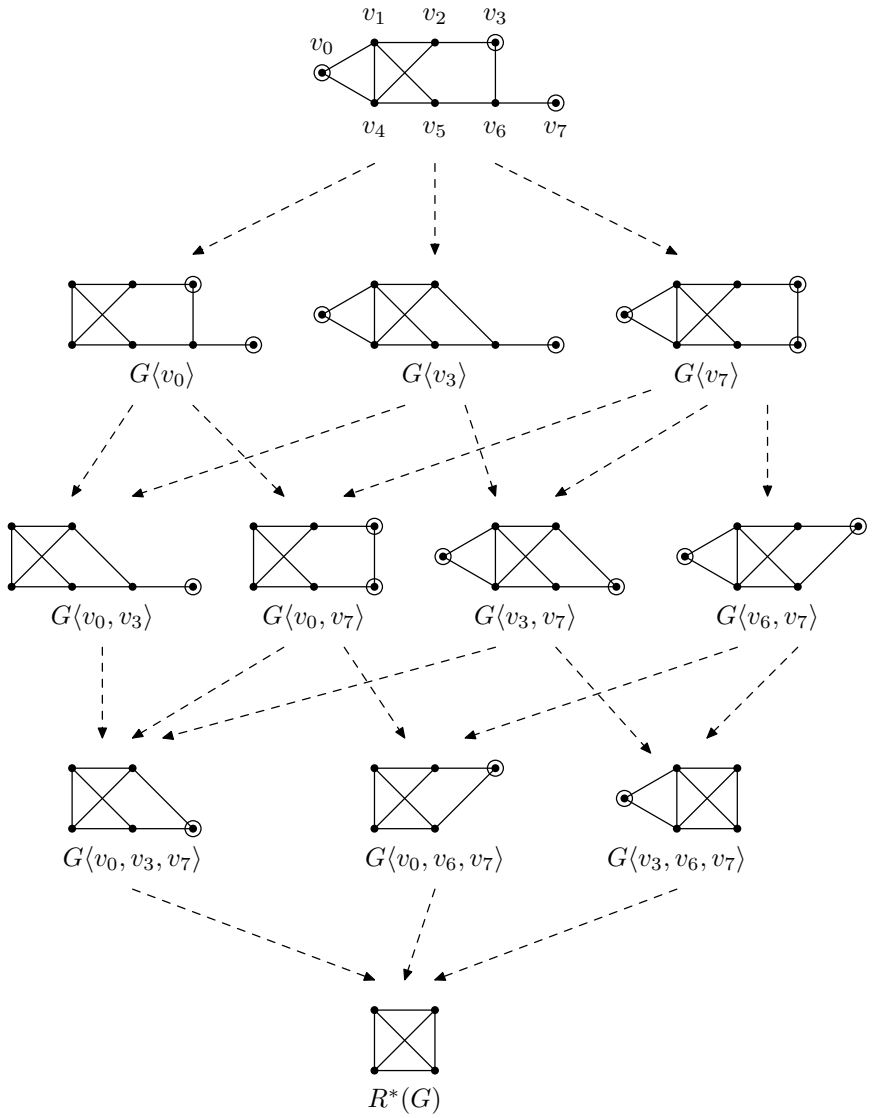


FIG. 3.1. *The above graphs illustrate the confluence of the reduction rules R_1 and R_2: No matter in which order nodes of degree one or two are reduced, the outcome is always the same. The rule R_0 cannot be applied because there are no isolated nodes, and we assume that $D = \emptyset$ for the sake of readability.*

does not change the degree of the other node or decreases it by one. That is, the reduction of x or y cannot render the reduction of the respective other node impossible. Hence, (x, y) and (y, x) are valid reduction sequences.

If none of the nodes is reduced according to R_2, it is also easy to see that $G\langle x, y\rangle = G\langle y, x\rangle$ because the resulting graph is $G \setminus \{x, y\}$. Otherwise, we may assume that x is reduced by R_2 without loss of generality. Since x and y are adjacent, y satisfies $\deg(y) \geq 1$. To see that the claimed equality holds in all remaining cases, check Figure 3.2. □

LEMMA 3.4. *Let $G = (V, E)$ be a graph, let $D \subseteq V$, and let $(v_1, \ldots, v_r)$ be a valid*

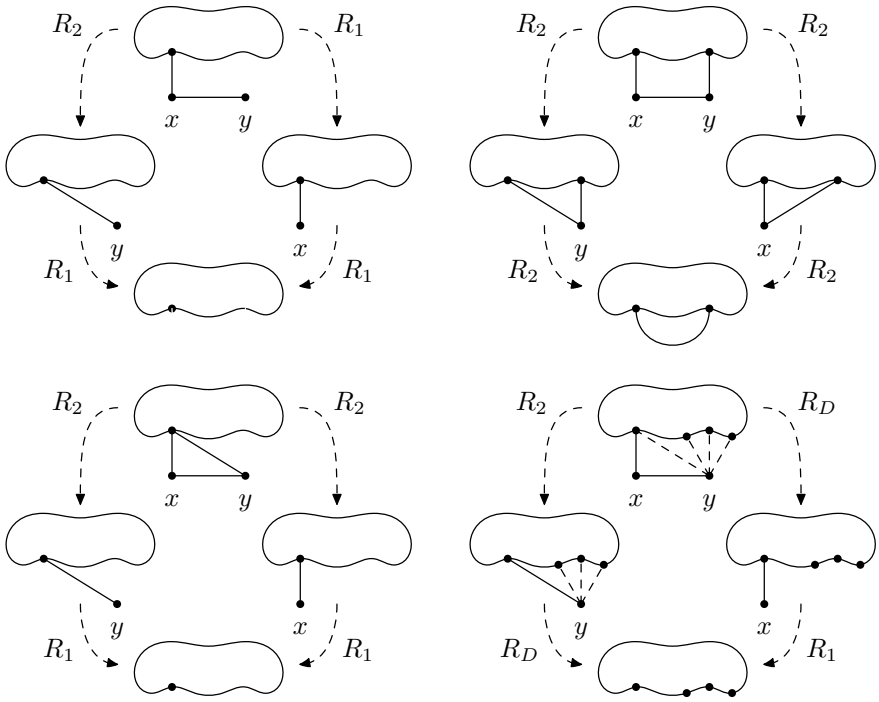


FIG. 3.2. *$G\langle x, y\rangle = G\langle y, x\rangle$ when x is reduced according to R_2.*

reduction sequence for G such that v_r is also reducible in G. Then $(v_r, v_1, \ldots, v_{r-1})$ is also a valid reduction sequence for G, and

$$G\langle v_1, \ldots, v_r\rangle = G\langle v_r, v_1, \ldots, v_{r-1}\rangle.$$

Proof. The claim is obvious for $r \leq 1$ and is given by Lemma 3.3 for $r = 2$. We show the cases $r > 2$ by induction on r.

Recall that $(v_1, \ldots, v_r)$ is a valid reduction sequence for G, and therefore (v_{r-1}, v_r) is a valid reduction sequence for $G\langle v_1, \ldots, v_{r-2}\rangle$. In particular, v_{r-1} is reducible in $G\langle v_1, \ldots, v_{r-2}\rangle$. Since v_r is also reducible in G, it must be reducible in $G\langle v_1, \ldots, v_{r-2}\rangle$ as well.

Lemma 3.3 guarantees that (v_r, v_{r-1}) and (v_{r-1}, v_r) are valid reduction sequences for $G\langle v_1, \ldots, v_{r-2}\rangle$ and that

$$G\langle v_1, \ldots, v_{r-2}\rangle\langle v_{r-1}, v_r\rangle = G\langle v_1, \ldots, v_{r-2}\rangle\langle v_r, v_{r-1}\rangle.$$

For the original valid reduction sequence, we find that

$$\begin{aligned} G\langle v_1, \ldots, v_r\rangle &= G\langle v_1, \ldots, v_{r-2}\rangle\langle v_{r-1}, v_r\rangle \\ &= G\langle v_1, \ldots, v_{r-2}\rangle\langle v_r, v_{r-1}\rangle = G\langle v_1, \ldots, v_{r-2}, v_r\rangle\langle v_{r-1}\rangle. \end{aligned}$$

This shows that $(v_1, \ldots, v_{r-2}, v_r, v_{r-1})$ is valid for G, and $(v_1, \ldots, v_{r-2}, v_r)$ is valid for G as well because it is a prefix.

Owing to the induction hypothesis, we know that $(v_r, v_1, \ldots, v_{r-2})$ is valid for G, too, and

$$G\langle v_1, \ldots, v_{r-2}, v_r\rangle = G\langle v_r, v_1, \ldots, v_{r-2}\rangle.$$

We get

$$\begin{aligned} G\langle v_1,\dots,v_r\rangle &= G\langle v_1,\dots,v_{r-2}\rangle\langle v_{r-1},v_r\rangle \\ &= G\langle v_1,\dots,v_{r-2}\rangle\langle v_r,v_{r-1}\rangle \\ &= G\langle v_1,\dots,v_{r-2},v_r\rangle\langle v_{r-1}\rangle \\ &= G\langle v_r,v_1,\dots,v_{r-2}\rangle\langle v_{r-1}\rangle \\ &= G\langle v_r,v_1,\dots,v_{r-2},v_{r-1}\rangle \end{aligned}$$

by putting all the pieces together. ∎

LEMMA 3.5. *Let $G=(V,E)$ be a graph, let $D\subseteq V$, and let $(v_1,\dots,v_r)$ as well as $(v_{\pi(1)},\dots,v_{\pi(r)})$ be two valid reduction sequences for G, where $\pi\in S_r$ is some permutation. Then*

$$G\langle v_1,\dots,v_r\rangle = G\langle v_{\pi(1)},\dots,v_{\pi(r)}\rangle.$$

Proof. The claim is obvious for $r\leq 1$ and is given by Lemma 3.3 for $r=2$. We show the cases $r>2$ by induction on r.

If $\pi(1)=1$, we can apply the induction hypothesis directly to see that

$$\begin{aligned} G\langle v_1,\dots,v_r\rangle &= G\langle v_1\rangle\langle v_2,\dots,v_r\rangle \\ &= G\langle v_1\rangle\langle v_{\pi(2)},\dots,v_{\pi(r)}\rangle \\ &= G\langle v_{\pi(1)},\dots,v_{\pi(r)}\rangle. \end{aligned}$$

Otherwise, let σ denote the sequence obtained from $(v_{\pi(1)},\dots,v_{\pi(r)})$ by removing v_1. Lemma 3.4 guarantees that

$$G\langle v_{\pi(1)},\dots,v_{\pi(r)}\rangle = G\langle v_1\rangle\langle\sigma\rangle.$$

Owing to the induction hypothesis, we get

$$G\langle v_1,\dots,v_r\rangle = G\langle v_1\rangle\langle v_2,\dots,v_r\rangle = G\langle v_1\rangle\langle\sigma\rangle,$$

and this entails the claim. ∎

LEMMA 3.6. *Let $G=(V,E)$ be a graph, let $D\subseteq V$, and let $(u_1,\dots,u_r)$ as well as $(v_1,\dots,v_s)$ be two valid reduction sequences for G such that $\{u_1,\dots,u_r\}$ and $\{v_1,\dots,v_s\}$ are disjoint. Then $(u_1,\dots,u_r,v_1,\dots,v_s)$ and $(v_1,\dots,v_s,u_1,\dots,u_r)$ are valid reduction sequences for G as well.*

Proof. Note that for any reducible $w\in V$, the reduced graph $G\langle w\rangle$ still contains all nodes from G with the sole exception of w. Consequently, $G\langle u_1,\dots,u_r\rangle$ still contains $v_1,\dots,v_s$. Note also that the degree of all nodes in $G\langle w\rangle$ is smaller than or the same as that in G, implying that v_1 is reducible in $G\langle u_1,\dots,u_r\rangle$.

Due to these facts, $(u_1,\dots,u_r,v_1)$ is a valid reduction sequence for G. Since v_2 is reducible in $G\langle v_1\rangle$ and $G\langle u_1,\dots,u_r,v_1\rangle = G\langle v_1,u_1,\dots,u_r\rangle$ according to Lemma 3.5, we know that v_2 is also reducible in $G\langle u_1,\dots,u_r,v_1\rangle$. In particular, (v_2) is a valid reduction sequence for $G\langle u_1,\dots,u_r,v_1\rangle$, and $(u_1,\dots,u_r,v_1,v_2)$ is a valid reduction sequence for G. Continuing in the same way, we can see that $(u_1,\dots,u_r,v_1,\dots,v_s)$ is a valid reduction sequence for G. Analogously, this statement holds for the sequence $(v_1,\dots,v_s,u_1,\dots,u_r)$. ∎

LEMMA 3.7. *Let $G=(V,E)$ be a graph, let $D\subseteq V$, and let $\sigma_1\tau_1$, $\sigma_2\tau_2$ be two valid reduction sequences for G. Furthermore, assume that $\sigma_1\cap\sigma_2\tau_2=\emptyset$ and $\sigma_2\cap\sigma_1\tau_1=\emptyset$.*

Then there are sequences μ_1 and μ_2 such that $\sigma_1\tau_1\sigma_2\mu_1$ and $\sigma_2\tau_2\sigma_1\mu_2$ are valid reduction sequences for G with $G\langle\sigma_1\tau_1\sigma_2\mu_1\rangle = G\langle\sigma_2\tau_2\sigma_1\mu_2\rangle$.

Proof. We use induction on $|\sigma_1\tau_1|+|\sigma_2\tau_2|$. Let us first assume that $|\sigma_1|+|\sigma_2| > 0$.

Since σ_1 and σ_2 are disjoint, Lemma 3.6 implies that $\sigma_1\sigma_2$ and $\sigma_2\sigma_1$ are valid reduction sequences for G. Lemma 3.5 then implies that

$$G\langle\sigma_1\sigma_2\rangle = G\langle\sigma_2\sigma_1\rangle. \tag{3.1}$$

We now know that $\sigma_2\tau_2$ and $\sigma_2\sigma_1$ are both valid reduction sequences for G. Hence, τ_2 and σ_1 are valid reduction sequences for $G\langle\sigma_2\rangle$. Furthermore, they are disjoint. Again, Lemmas 3.6 and 3.5 imply that $\sigma_1\tau_2$ and $\tau_2\sigma_1$ are valid reduction sequences for $G\langle\sigma_2\rangle$ and that $G\langle\sigma_2\rangle\langle\sigma_1\tau_2\rangle = G\langle\sigma_2\rangle\langle\tau_2\sigma_1\rangle$. In the same way we get $G\langle\sigma_1\rangle\langle\sigma_2\tau_1\rangle = G\langle\sigma_1\rangle\langle\tau_1\sigma_2\rangle$. We can rewrite these two equalities as

$$G\langle\sigma_2\sigma_1\tau_2\rangle = G\langle\sigma_2\tau_2\sigma_1\rangle \quad \text{and} \quad G\langle\sigma_1\sigma_2\tau_1\rangle = G\langle\sigma_1\tau_1\sigma_2\rangle. \tag{3.2}$$

Let $G' = G\langle\sigma_1\sigma_2\rangle$, $\sigma_1' = \sigma_2' = \varepsilon$, $\tau_1' = \tau_1$, and $\tau_2' = \tau_2$. Note that $\sigma_1'\tau_1'$ and $\sigma_2'\tau_2'$ are both valid for G' due to (3.2). Of course, $\sigma_1' \cap \sigma_2'\tau_2' = \sigma_2' \cap \sigma_1'\tau_1' = \emptyset$ because σ_1' and σ_2' are empty. All preconditions of this lemma are fulfilled, and $|\sigma_1'\tau_1'| + |\sigma_2'\tau_2'| < |\sigma_1\tau_1| + |\sigma_2\tau_2|$. We can thus use the induction hypothesis to show the existence of μ_1' and μ_2' such that $\tau_1'\mu_1'$ and $\tau_2'\mu_2'$ are valid reduction sequences for G' and that $G'\langle\tau_1'\mu_1'\rangle = G'\langle\tau_2'\mu_2'\rangle$. If we choose $\mu_1 = \mu_1'$ and $\mu_2 = \mu_2'$, then this is exactly the same as

$$G\langle\sigma_1\sigma_2\tau_1\mu_1\rangle = G\langle\sigma_1\sigma_2\tau_2\mu_2\rangle. \tag{3.3}$$

Using all of the above we get (see Figure 3.3 for an illustration)

$$G\langle\sigma_1\tau_1\sigma_2\mu_1\rangle \overset{(3.2)}{=} G\langle\sigma_1\sigma_2\tau_1\mu_1\rangle \overset{(3.3)}{=} G\langle\sigma_1\sigma_2\tau_2\mu_2\rangle \overset{(3.1)}{=} G\langle\sigma_2\sigma_1\tau_2\mu_2\rangle \overset{(3.2)}{=} G\langle\sigma_2\tau_2\sigma_1\mu_2\rangle.$$

The other case is that $\sigma_1 = \sigma_2 = \varepsilon$. If $\tau_1 = \varepsilon$ as well, the statement of the lemma holds because setting $\mu_1 = \tau_2$ and $\mu_2 = \varepsilon$ guarantees that $\sigma_1\tau_1\sigma_2\mu_1 = \sigma_2\tau_2\sigma_1\mu_2$ and thus $G\langle\sigma_1\tau_1\sigma_2\mu_1\rangle = G\langle\sigma_2\tau_2\sigma_1\mu_2\rangle$. Otherwise, if τ_1 is not empty, we may furthermore assume that the first vertex in τ_1 also occurs in τ_2: if it did not, we could shift the first node of τ_1 into σ_1 and apply the argument from the above first case.

Now define $\tau_1 = v\tau_1'$ and $\tau_2 = \tau_2' v \tau_2''$. Applying Lemma 3.4 to $\tau_2' v$ yields $G\langle\tau_2' v\rangle = G\langle v\tau_2'\rangle$. This entails $G\langle v\tau_2'\tau_2''\rangle = G\langle\tau_2\rangle$. Furthermore, both τ_1' and $\tau_2'\tau_2''$ are valid reduction sequences for $G\langle v\rangle$. Owing to the induction hypothesis with respect to $G\langle v\rangle$ and the sequences below, there are μ_1 and μ_2 such that

$$G\langle\tau_1\mu_1\rangle = G\langle v\rangle\langle\tau_1'\mu_1\rangle \overset{\text{i.h.}}{=} G\langle v\rangle\langle\tau_2'\tau_2''\mu_2\rangle = G\langle\tau_2\mu_2\rangle,$$

which completes the proof. ∎

THEOREM 3.8. *Let $G = (V, E)$ be a graph and $D \subseteq V$. Then $R^*(G)$ is well defined; i.e., if τ_1 and τ_2 are two valid reduction sequences for G of maximal length, then $G\langle\tau_1\rangle = G\langle\tau_2\rangle$.*

Proof. Let $\sigma_1 = \sigma_2 = \varepsilon$; then G, D, $\sigma_1\tau_1$, and $\sigma_2\tau_2$ satisfy the conditions of Lemma 3.7. Thus there are sequences μ_1 and μ_2 such that $\sigma_1\tau_1\sigma_2\mu_1 = \tau_1\mu_1$ and $\sigma_2\tau_2\sigma_1\mu_2 = \tau_2\mu_2$ are valid reduction sequences for G with $G\langle\tau_1\mu_1\rangle = G\langle\tau_2\mu_2\rangle$. The fact that $\tau_1\mu_1$ is a valid reduction sequence for G and that τ_1 is a reduction sequence of maximal length implies $\mu_1 = \varepsilon$. Using the same argument, we obtain $\mu_2 = \varepsilon$. Hence, $G\langle\tau_1\rangle = G\langle\tau_2\rangle$. ∎

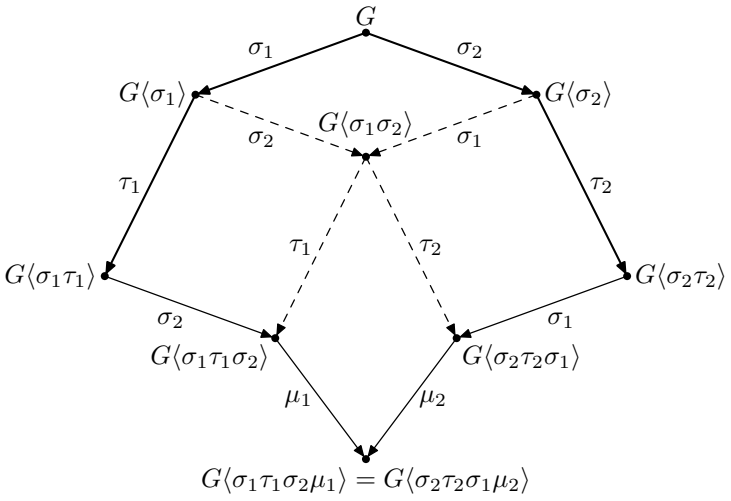


FIG. 3.3. *Confluent sequences.*

4. Bounds on treewidth and pathwidth. Now that we have established the confluence of our reduction rules, we continue by investigating their influence on the treewidth of graphs. The following lemmas reveal two important properties: graphs of treewidth at most two collapse upon application of the reduction rules, whereas the rules R_0, R_1, and R_2 cannot affect any treewidth greater than two. These lemmas constitute important building blocks for the main result of this section, namely, the bound of $m/5.769 + O(\log n)$ on the pathwidth of graphs.

LEMMA 4.1. *Let $G = (V, E)$ be a graph, $D \subseteq V$, and $tw(G) \leq 2$. Then $R^*(G)$ is empty.*

Proof. Connected graphs with treewidth at most two are reduced by R_1 and R_2 to a single node [2]. □

LEMMA 4.2. *Let $G = (V, E)$ be a connected graph, let $tw(G) > 2$, and let G' be a graph obtained from G by applying R_0, R_1, or R_2. Then $tw(G) = tw(G')$.*

Proof. Note that the reduction rules can be inverted easily. The inverse of R_0 or R_1 is to add or connect a new node to the graph, and the inverse of R_2 is to subdivide an edge (with or without keeping a copy of that edge). We begin with a tree decomposition for G' and construct a tree decomposition for G.

Let us first investigate the case that R_0 or R_1 was applied to turn G into G'. Adding an isolated node to a tree decomposition is trivial, and in order to connect a new node v to some node w in G', it suffices to find a bag B with $w \in B$ and attach a new bag $B' = \{v, w\}$ to B.

Otherwise, R_2 was applied. To subdivide an edge $\{w_1, w_2\}$ by a new node v (with or without keeping a copy of that edge), it suffices to find a bag B with $w_1, w_2 \in B$ and attach a new bag $B' = \{v, w_1, w_2\}$ to B.

In either case, the resulting tree of bags is a tree decomposition for G. Moreover, we added only a bag B' of size at most three. If $tw(G') \leq 2$, then $tw(G) \leq 2$, which contradicts the assumption that $tw(G) \geq 3$ from the statement of this lemma. Otherwise, we have $tw(G') \geq 3$ and $tw(G') = tw(G)$ because the old tree decomposition already contains a bag at least as large as B'. □

LEMMA 4.3. *Let $G = (V, E)$ be a graph such that $R^*(G) = \emptyset$ where $D = \emptyset$. Then $pw(G) = O(\log |V|)$.*

Proof. Lemma 4.2 implies $tw(G) \leq 2$; thus there is a separator U of size three such that all components of $G \setminus U$ have a size of at most $(n-2)/2$ [3]. Recursively construct path decompositions for these components, concatenate them, and add U to each bag to get a path decomposition for G. The recursion depth is logarithmic. ∎

In order to bound the treewidth of a graph by $m/5.769 + O(\log n)$, we present a construction of tree decompositions based on the iterated removal of nodes where reduction rules are applied whenever possible. If the graph splits into several components, the construction can be performed for each of the components independently.

Clearly, the removal of a node—combined with subsequent reductions—leads to a loss of edges as well. Since the number of edges that vanish upon deletion of a node is small in a few cases, but larger on average, we employ an amortized analysis using node potentials. To do this, we first define a potential function $\phi\colon V \to \mathbf{Q}$ as follows.

DEFINITION 4.4.

$$\phi(v) := \begin{cases} 0 & \text{if } \deg(v) < 3, \\ 25/26 & \text{if } \deg(v) = 3, \\ 25/13 & \text{if } \deg(v) = 4, \\ \deg(v)/2 & \text{if } \deg(v) \geq 5. \end{cases}$$

The continuation Φ maps entire graphs to the sum of the potentials over all nodes: $\Phi(G) := \sum_{v \in V} \phi(v)$. Observe that this sum—the potential of a graph—never exceeds the number of edges in any graph.

LEMMA 4.5. *Let $G = (V, E)$ be a reduced connected graph that contains a node of degree at least four and is not five-regular. There is a node of maximum degree whose removal decreases the potential of G by at least $75/13 > 5.769$.*

Proof. Choose a node v of maximum degree and recall that $\deg(v) \geq 4$. In the special case that the maximum degree equals five, choose a node v of degree five at least one of whose neighbors has lower degree. Observe also that every node in G has degree at least three because G is assumed to be a reduced graph.

The removal of v decreases the potential in two ways. First, the deletion of v lowers the potential of G by $\phi(v)$. Second, the deletion of v lowers the degree of each neighbor of v by one, leading to another loss in potential. If such a neighbor has degree three or four, its potential decreases by $25/26$. If its degree equals five, its potential decreases by only $5/2 - 25/13 = 15/26$. In all other cases, the potential of the neighbor decreases by $1/2$.

We employ a case distinction to verify that the potential of G is lowered by at least $75/13$ when v is removed.

If $\deg(v) = 4$, each neighbor of v has degree three or four, and the above considerations imply that the total loss of potential amounts to exactly $25/13 + 4 \cdot 25/26 = 75/13$.

In the special case that $\deg(v) = 5$, at least one neighbor of v has degree three or four as detailed above. The removal of v thus leads to a total loss of potential of at least $5/2 + 4 \cdot 15/26 + 25/26 = 75/13$.

If $\deg(v) \geq 6$, the potential drops by at least $6/2 + 6 \cdot 1/2 = 6$. ∎

An exceedingly helpful case is reflected by the following lemma.

LEMMA 4.6. *Let $G = (V, E)$ be a connected graph of maximum degree five that has been reduced according to the rules R_0, R_1, and R_2. Let $v \in V$ be a node with $\deg(v) = 5$ such that $R^*(G \setminus \{v\})$ contains a five-regular component C. Then $\Phi(C) \leq \Phi(G) - 95/13$.*

Proof. Since G is connected, v has a neighbor of degree three. Otherwise, each component of $G \setminus \{v\}$ would contain at least one node of degree three or four and no reduction rules could be applied. This contradicts the existence of a five-regular component in $R^*(G \setminus \{v\})$.

The most simple way to obtain a five-regular component is to remove a node, all of whose neighbors are of degree three. After the removal all neighbors are contracted by R_2, leading to a possibly five-regular component. Regardless of the structure of the resulting graph, the potential $\Phi(G)$ decreases by at least $5/2 + 5 \cdot 25/26 = 95/13$.

In any other case resulting in a five-regular component, the removal of v decreases the degree of some of its neighbors to three or four. In order to obtain a five-regular component, these neighbors must either be part of a different component or be reduced by some further reduction rules. Let n_i denote the number of nodes in $N(v)$ with degree i. Then, on removal of v the potential of G decreases by only $5/2 + n_3 \cdot 25/26 + n_4 \cdot (25/13 - 25/26) + n_5 \cdot (5/2 - 25/13)$, but the potential of C can be bounded by

$$\Phi(G) - 5/2 - n_3 \cdot 25/26 - n_4 \cdot 25/13 - n_5 \cdot 5/2.$$

Therefore, the potential of C is at most $\Phi(G) - 95/13$. ∎

The above lemmas enable us to prove the main result of this paper.

THEOREM 4.7. *Let $G = (V, E)$ be a graph. Then $tw(G) \leq |E|/5.769 + O(\log n)$, and a respective tree decomposition can be obtained in polynomial time.*

Proof. Without loss of generality, assume that G is connected. We prove the claim constructively by giving an algorithm that outputs the respective tree decomposition. Basically, the algorithm is really simple: it keeps on removing nodes of maximum degree and adds them to each bag of the tree decomposition. After each node removal, the aforementioned reduction rules are applied immediately. As soon as the graph becomes cubic, we obtain the rest of the tree decomposition using a technique by Fomin and Høie.

The proof requires us to deal with some technicalities in order to obtain the desired result. First, the removal of a node may split the graph into several components, but these can be handled independently. Second, we avoid removing a degree-five node with only degree-five neighbors whenever possible; this is a critical case in the analysis.

To see how the algorithm can be employed to construct a tree decomposition, let $G = (V, E)$ denote the currently inspected graph and v the node selected for removal. If $G' = G \setminus \{v\}$ is connected, a tree decomposition for G can be obtained by adding v to each bag of a tree decomposition for G'. In particular, this operation cannot invalidate the tree decomposition. Otherwise, if G' consists of several components, a tree decomposition for G can be obtained as follows: after adding v to each bag in the tree decompositions of the components, connect a new bag $\{v\}$ to an arbitrary bag of every such decomposition. Again, it is easy to verify that the resulting tree of bags is a tree decomposition.

As detailed in Lemmas 4.1 and 4.2, the reduction rules do not increase the potential or the treewidth of a graph. Moreover, as described in the proof of Lemma 4.2, the tree decomposition can be updated accordingly whenever a reduction rule has been applied. It remains only to show that the size of the bags does not exceed the claimed bound, which is done using an amortized analysis using node potentials.

We distinguish three phases. As long as the graph contains nodes of degree at least six, we are in the first phase. While the maximum degree equals five or four, we are in the second phase. The third phase begins as soon as the maximum degree decreases to three or less. Observe that the maximum degree, as well as the degree

Algorithm T
Input: A reduced graph G
Output: A tree decomposition for G

```
(1)  B := ∅;
(2)  if G has maximum degree at most three then
       return path decomposition as computed by Fomin and Høie;
(3)  if G consists of several independent components G_1, ..., G_l
(4)       then connect B to one bag from each T(G_i);
(5)       return this tree decomposition;
(6)  else
(7)       choose a preferable node v;
(8)       G' = G \ {v}; T' = T(R*(G'));
(9)       Update T' according to every applied reduction rule;
(10)      Add v to each bag of T';
(11)      return T';
```

FIG. 4.1. *Algorithm T computes a tree decomposition. If G is five-regular, then every node is preferable. Otherwise, every node of maximum degree is preferable unless it has exactly five neighbors, each of which have degree five.*

of every node in the graph, decreases monotonically as we proceed to remove nodes, implying that the phases are traversed in the given order.

Within the first phase, each step decreases the potential by at least $3+6\cdot 1/2 = 6$: a node of degree at least six has a potential of at least 3, and each of its neighbors loses an edge, which decreases the potential by at least $1/2$ per neighbor.

Whenever the removed node disconnects the graph, it suffices to compute the tree decomposition for each of the respective components independently. At any point, we may thus restrict our analysis to the component having the largest potential. For the second phase, it hence suffices to analyze the cases in which the graph is connected and has maximum degree four or five.

According to Lemma 4.5, each step in the second phase decreases the potential by $75/13$ unless the graph is five-regular. When removing a node from a five-regular graph, the potential decreases by only $5/2 + 5 \cdot (5/2 - 25/13) = 70/13$. However, Lemma 4.6 implies that in the last step the potential has been decreased by at least $95/13$, except for the very first step of this phase, whose constant additional cost is hidden in the $O(\log n)$ term. Thus, the average loss is $165/26 > 75/13$.

As soon as we enter the third phase, the remaining graph $G' = (V', E')$ is either three-regular or empty (due to reductions). It obviously suffices to consider the three-regular case, in which $|V'| = \Phi(G') \cdot 26/25$. According to a result by Fomin and Høie [5], the pathwidth of an n-node cubic graph is bounded by $(1+\varepsilon)n/6 + O(\log n)$, where $\varepsilon > 0$ is an arbitrarily small constant. This implies a bound of

$$(1+\varepsilon)(\Phi(G') \cdot 13/75) + O(\log \Phi(G')) \leq \Phi(G')/5.769 + O(\log |V'|)$$

on the treewidth of G'. A respective tree decomposition can be computed in polynomial time [5]. □

A tree-decomposition algorithm that employs the construction used in Theorem 4.7 is depicted in Figure 4.1.

LEMMA 4.8. *Let G be a graph, $D = \emptyset$, and $(V', E') = R^*(G)$. Then every (connected) component of $G \setminus V'$ is connected to at most two vertices in $G[V']$.*

Proof. Let $C = G[\{v_1, \ldots, v_r\}]$ be a (connected) component of $G \setminus V'$. G has been reduced to (V', E') by a valid reduction sequence σ. Observe that σ contains $v_1, \ldots, v_r$ without loss of generality in this order. Moreover, v_1 is reducible in G because no neighbor of v_1 is removed before v_1. Lemma 3.4 shows that we can move v_1 to the front of σ. Repeating this argument, we see that $(v_1, \ldots, v_r)$ is a valid reduction sequence for G. Moreover, applying any reduction rule on C does not affect the connectivity of the remaining nodes in C, since only nodes of degree one or less are removed and nodes of degree two are contracted.

Let

$$V_i := \{\, v \in V' \mid v \text{ is a neighbor of } \{v_i, \ldots, v_r\} \text{ in } G\langle v_1, \ldots, v_{i-1}\rangle \,\}.$$

We claim that $V_i = V_{i+1}$ for $1 \le i \le r-1$.

If v_i has no neighbor in V_i (in the graph $G\langle v_1, \ldots, v_{i-1}\rangle$), the claim obviously holds. Otherwise, v_i is of degree two with neighbors $u \in V'$ and $w \in \{v_{i+1}, \ldots, v_r\}$. $V_i \setminus \{u\} \subseteq V_{i+1}$, as all these nodes have neighbors in $\{v_{i+1}, \ldots, v_r\}$.

Applying R_2 on v_i adds a new edge between u and w. Since $w \in G\langle v_1, \ldots, v_i\rangle$, $u \in V_{i+1}$. Thus $V_1 = V_r$.

Now, if $|V_r| \ge 3$, then v_r is not reducible in $G\langle v_1, \ldots, v_{r-1}\rangle$, which is a contradiction. ∎

THEOREM 4.9. *Let $G = (V, E)$ be a graph. Then $pw(G) \le |E|/5.769 + O(\log n)$, and a respective path decomposition can be obtained in polynomial time.*

Proof. Let D be the set of nodes that have been chosen as preferable nodes in line (7) of Algorithm T. The algorithm transforms G into a cubic graph $G\langle\sigma\rangle$, where σ is a valid reduction sequence with respect to D. Note that every node in D is reducible in G. By Lemma 3.4 there is a valid reduction sequence $\sigma' = (d_1, \ldots, d_r, v_1, \ldots, v_s)$ that is a permutation of σ and $d_i \in D$, $v_i \notin D$. Moreover, $G\langle\sigma\rangle = G\langle\sigma'\rangle$.

We will modify Algorithm T so as to construct a path decomposition instead of a tree decomposition.

Without loss of generality, we assume $G\langle\sigma'\rangle$ is connected. Otherwise, we apply the following argument for each component separately.

Let $P = (P_1, \ldots, P_t)$ be a path decomposition for $G\langle\sigma'\rangle$ as computed by Fomin and Høie. A component C of $G[\{v_1, \ldots, v_s\}]$ has at most two neighbors in $G\langle\sigma'\rangle$ according to Lemma 4.8. If there are indeed two neighbors, they must be connected by an edge in $G\langle\sigma'\rangle$ as the path connecting both neighbors in C has been contracted to a single edge. Thus the neighbors occur together in a bag of P. Let P_i be the smallest bag in P that contains all neighbors of C and $P' = (P'_1, \ldots, P'_k)$ be a path decomposition of C. Since C is series-parallel, the width of P' is only $O(\log n)$. Then $(P'_1 \cup P_i, \ldots, P'_k \cup P_i)$ is a path decomposition for $G[V(C) \cup P_i]$ and $(P_1, \ldots, P_i, P'_1 \cup P_i, \ldots, P'_k \cup P_i, P_{i+1}, \ldots P_t)$ is a path decomposition for $G[U]$, where U consists of all nodes in C and $G\langle\sigma'\rangle$. Notice that we can do this for all components of $G[\{v_1, \ldots, v_s\}]$ in parallel, such that the size of the resulting bags is still bounded by the width of P plus $O(\log n)$, since the original bags from P remain untouched and thus can be used as smallest bag P_i.

Therefore, we obtain path decompositions for every connected component of $G\langle d_1, \ldots, d_r\rangle$. To obtain path decompositions for each component of $G\langle d_1, \ldots, d_{r-1}\rangle$, we proceed as follows: We add d_r to every bag of the path decomposition of each

component that is adjacent to d_r just as in Algorithm T. Afterward, we connect these path decompositions as a path.

Compared to the tree decompositions described in Theorem 4.7, only the incorporation of the path decompositions of $G[\{v_1, \ldots, v_s\}]$ increases the size of the bags as described above. We obtain a bound of

$$O(\log n) + |E|/5.769 + O(\log n) = |E|/5.769 + O(\log n)$$

for the width of our path decomposition. □

Employing a framework for algorithms that work on tree decompositions by Telle and Proskurowski [20], one immediately obtains the following result.

COROLLARY 4.10. MAX-2SAT *and* MAX-CUT *can be solved in* $O^*(2^{m/5.769})$ *using exponential space.*

The exponential space complexity of the resulting algorithms is due to dynamic programming on the actual tree decomposition. As we will see in the next section, tree decompositions that were computed according to Theorem 4.7 have a unique structure: instead of adding nodes to the tree decomposition when they are removed from the graph in the first two phases and solving the problem at hand with the framework by Telle and Proskurowski, we can simply branch on these nodes if there are appropriate reduction rules for that problem. Since branching requires only polynomial space, this will enable us to get rid of the exponential space complexity.

Unfortunately, the algorithm by Fomin and Høie employed in the third phase does not necessarily output decompositions of the aforementioned structure. Since this forbids us to switch to algorithms that branch directly, the third phase forces us to use the Telle–Proskurowski approach and thus requires exponential space.

5. Algorithms for MAX-CUT and MAX-2SAT. In order to enforce polynomial space complexity and to solve the problem discussed at the end of the previous section, we now abandon the special processing of cubic graphs. The resulting algorithms for problems like MAX-CUT and MAX-2SAT solely rely on branching and avoid any dynamic programming. They guarantee polynomial space complexity at the expense of slightly worse runtime bounds. Again, branching is only possible for problems that can be represented as graph problems with appropriate reduction rules for nodes of degree at most two.

As a bonus, these branching-based algorithms are extremely intuitive. In contrast to previous algorithms [1, 14, 7], our algorithm for MAX-2SAT (see Figure 5.1) consists of only three reductions and straightforward branching. Whenever a variable x occurs with at most two other variables y, z, we can eliminate x by adding new clauses over y and z. If branching leads to several independent subformulas, we can solve these independently—a very natural reduction. Finally, the algorithm simply branches by setting a variable x to true or false, which is the most simple branching imaginable. The simple structure makes the resulting algorithms relatively efficient (the runtime bounds do not contain large hidden constants or polynomials) but also easy to implement and verify.

Since we cannot rely on the result for cubic graphs by Fomin and Høie [6] any longer, we need to redefine the node potentials. The following values turn out to be the best choice for our analysis.

Algorithm A
Input: A MAX-2SAT-formula F
Output: $A(F) = OptVal(F)$

(1) Reduce F by the reduction rules while possible;
(2) **if** $F = \{(k, \mathbf{T})\}$ **then return** k;
(3) **if** F consists of several independent subformulas $F_1, \ldots, F_l$
(4) **then return** $\sum_{i=1}^{l} A(F_i)$;
(5) **else**
(6) choose the preferable variable x;
(7) **return** $\max\{A(F[x]), A(F[\bar{x}])\}$;

FIG. 5.1. *A very simple algorithm for* MAX-2SAT *that does not use the connectivity graph directly.*

DEFINITION 5.1.

$$\psi(v) := \begin{cases} 0 & \text{if } \deg(v) < 3, \\ 30/23 & \text{if } \deg(v) = 3, \\ 45/23 & \text{if } \deg(v) = 4, \\ \deg(v)/2 & \text{if } \deg(v) \geq 5. \end{cases}$$

Again, Ψ maps entire graphs to the sum of the potentials over all nodes. As in Definition 4.4, the potential of any graph is bounded by its number of edges.

LEMMA 5.2. *Let $G = (V, E)$ be a connected graph of maximum degree four that has been reduced according to the rules R_0, R_1, and R_2. Let $v \in V$ be a node with $\deg(v) = 4$ such that $R^*(G \setminus \{v\})$ contains a four-regular component C. Then $\Phi(C) \leq \Phi(G) - 165/23$.*

Proof. This is can be proven analogously to Lemma 4.6. □

LEMMA 5.3. *Let $G = (V, E)$ be a graph such that $R^*(G)$ is not empty. If $R^*(G)$ consists of multiple components, then each component has a potential of at most $\Psi(G) - 120/23$.*

Proof. Each component of $R^*(G)$ contains a node v of degree at least three. The neighbors of v have degree at least three as well. Hence, the potential of each component is at least $4 \cdot 30/23$. If there are multiple components, the potential of each is thus bounded by $\Psi(G) - 120/23$. □

Using the above two lemmas, it is possible to find small node sets that either split a graph into several components of bounded potential or leave a trivial graph. This is formalized by the following theorem which is the backbone of the upcoming algorithms for MAX-2SAT and MAX-CUT.

THEOREM 5.4. *Let $G = (V, E)$ be a graph. There is a set $D \subseteq V$ such that either*

(i) *$R^*(G \setminus D)$ contains at least two components, each having a potential of at most $\Psi(G) - 5.217|D|$, or*

(ii) *$R^*(G \setminus D) = \emptyset$ and $|D| \leq \Psi(G)/5.217 + 1$.*

Proof. Let $G = (V, E)$ be a graph. If G has maximum degree at least five, removing a node of maximum degree and applying the reduction rules decreases the potential by at least $2.5 + 5 \cdot (2.5 - 45/23) = 120/23 > 5.217$. We may thus assume that G has maximum degree at most four.

Analogously to Theorem 4.7, we remove nodes of maximum degree until G either splits into several components or becomes empty. In doing so, we avoid nodes with

four neighbors of degree four if possible. Again, the potential decreases by at least 120/23 in any step, except for the aforementioned four-regular case. But even in this case with a loss of $45/23 + 4 \cdot 15/23 = 105/23$, there is an average loss of more than 120/23 according to Lemma 5.2, since the step before yields a loss of at least 165/23.

There is only one exception that does not allow for the above bonus argument, namely, the case when the graph is four-regular for the first time. Note that the loss of potential in this case amounts to only 105/23, which is 15/23 short of the desired value.

If we end up with an empty graph, the additional node in D is absorbed by the last summand in the bound $\Psi(G)/5.217 + 1$. Otherwise, if the graph breaks down into several components, the remaining potential is at most $\Psi(G) - 5.217|D| + 15/23$. Lemma 5.3 implies that each component has potential at most $\Psi(G) - 5.217|D| + 15/23 - 120/23 < \Psi(G) - 5.217|D|$.

Note that according to our results from section 3, reducing the graph $G \setminus D$ yields exactly the same graph as removing the nodes in D successively and reducing the remaining graph in each step. Hence, the nodes selected by the above algorithm constitute a set D with the desired properties. □

Note that, similar to the proofs from section 4, this result can be used to bound the pathwidth of sparse graphs by $m/5.127 + 3$. While this bound is worse than the one obtained earlier, the corresponding path decompositions have nice properties that can be exploited in direct algorithms, as we will see shortly.

We now possess the graph-theoretical means to construct the desired polynomial-space algorithms. When F is a MAX-2SAT formula, we use G_F to denote the connectivity graph of F: each variable in F is represented by a node in G_F, and two nodes are connected if and only if the formula contains a clause consisting of the two corresponding variables. Observe that the connectivity graph does not represent negations or weights in the formula. For instance, $f = (x_1 \vee x_2) \wedge (x_1 \vee x_2) \wedge (x_2 \vee x_3)$ and $g = (\overline{x_1} \vee x_2) \wedge (\overline{x_2} \vee x_3)$ have identical connectivity graphs G_f and G_g. As a consequence, the formula F cannot be reconstructed from G_F.

In order to fix a terminology for the discussion of satisfiability problems, we adhere to the notation for weighted boolean formulas used by Gramm et al. [7].

DEFINITION 5.5. *A (weighted) clause is a pair (ω, S) where ω is an integer and S is a nonempty finite set of literals that does not contain, simultaneously, any variable together with its negation.*

A formula *F is a set of clauses, such that each set of literals appears in at most one clause.*

We call ω the weight *of a clause (ω, S) and define*

$$w_F(S) = \begin{cases} \omega & \text{if } (\omega, S) \in F, \\ 0 & \text{otherwise.} \end{cases}$$

In addition to usual clauses, we allow a special *true clause* $(\omega, \mathbf{T})$ which is satisfied by every assignment. (We also call it a $\mathbf{T}$-*clause*.) The operators + and − are defined as follows.

DEFINITION 5.6.

$$F + G = \{\, (w_F(S) + w_G(S), S) \mid w_F(S) + w_G(S) \neq 0 \,\},$$
$$F - G = \{\, (w_F(S) - w_G(S), S) \mid w_F(S) - w_G(S) \neq 0 \,\}.$$

For a literal l and a formula F, the formula $F[l]$ is obtained by setting the value of l to $True$. For a set of literals $X = \{x_1, \ldots, x_k\}$, we set $F[X] := F[x_1][x_2]\ldots[x_k]$. For example,

$$\begin{aligned} F &= \{(5, \{x, y, \bar{z}\}), (3, \{y, \bar{z}\}), (2, \{\bar{x}, y\}), (-1, \{\bar{x}, y, \bar{z}\})\}, \\ F[x] &= \{(5, \mathbf{T}), (2, \{y, \bar{z}\}), (2, \{y\})\}, \\ F[\bar{x}] &= \{(1, \mathbf{T}), (8, \{y, \bar{z}\})\}, \\ F[x, y] &= \{(9, \mathbf{T})\}. \end{aligned}$$

DEFINITION 5.7. *The optimal value of a maximum weight assignment for formula F is defined as $OptVal(F) = \max_A\{\, \omega \mid (\omega, \mathbf{T}) \in F[A]\,\}$, where A is taken over all possible assignments.*

An assignment A is optimal *if $F[A]$ contains only one clause $(\omega, \mathbf{T})$ (or if it does not contain any clause, in this case we set $\omega = 0$) and $OptVal(F) = \omega$ $(= OptVal(F[A]))$. We call F and G max-equivalent if $OptVal(F) = OptVal(G)$.*

In order to design a branching algorithm for MAX-2SAT as suggested above (i.e., an algorithm that branches on nodes from the set D as constructed in Theorem 5.4), we must find reduction rules for formulas that correspond to removing nodes of degree at most one and contracting nodes of degree two. Clearly, a simple reduction rule suffices to remove nodes of degree zero from G_F: if a variable occurs only in unary clauses, it is optimal to choose the assignment that satisfies most of these clauses.

But what do we need to do with F in order to remove a degree-one node or contract a degree-two node in G_F? It is easy to see that we have to eliminate a variable x that occurs with exactly one or two other variables, respectively. In order to maintain a max-equivalent formula, these steps require us to introduce new clauses.

Since we branch on nodes in the connectivity graph that hides negations it is straightforward to analyze the running time without respect to negations as well. This notion is reflected by the following definition.

DEFINITION 5.8. *Let F be a SAT formula. For each clause C, we call the set of variables that occur in C the* clause type*. The* clause types *of F are the clause types of the clauses in F.*

For example, $(x_1 \vee x_2)$ and $(\overline{x_1} \vee x_2)$ have the same type and may thus be counted as one entity in the analysis. This way of measuring the complexity of a formula by the number of clause types has a crucial advantage: we may employ reduction rules that introduce additional clauses of existing types without raising the potential.

DEFINITION 5.9. *Let F be a 2SAT formula. We call the variable x a* companion *(*of *y) if there is a unique variable $y \neq x$ that occurs together with x in a clause.*

In terms of the respective connectivity graph G_F, the variable x is a companion if and only if the degree of x in G_F is one.

LEMMA 5.10 (companion reduction rule). *Let F be a 2SAT formula. If x is a companion, we can transform F into a max-equivalent formula F' containing the same variables except for x, where $G_{F'} = G_F \setminus \{x\}$. This can be done in polynomial time.*

Proof. Let F be a formula, let x be a companion of y, let F' consist of all clauses in F with an occurrence of the variable x, and let $F'' = F \setminus F'$. Let, furthermore, $a = OptVal(F'[y])$, $b = OptVal(F'[\bar{y}])$, and

$$H = \begin{cases} \{(b, \mathbf{T}), (a - b, \{y\})\} & \text{if } a > b, \\ \{(a, \mathbf{T}), (b - a, \{\bar{y}\})\} & \text{otherwise.} \end{cases}$$

It is easy to see that $a = OptVal(H[y])$ and $b = OptVal(H[\bar{y}])$. We immediately get

$$\begin{aligned} &OptVal(H + F'') \\ &\quad = \max\{OptVal(H[y]) + OptVal(F''[y]), OptVal(H[\bar{y}]) + OptVal(F''[\bar{y}])\} \\ &\quad = \max\{OptVal(F'[y]) + OptVal(F''[y]), OptVal(F'[\bar{y}]) + OptVal(F''[\bar{y}])\} \\ &\quad = OptVal(F' + F'') = OptVal(F). \end{aligned}$$

Hence, we can replace F by the max-equivalent formula $H + F''$. Note that it is easy to calculate a and b and that $H + F''$ does not contain the variable x anymore. ∎

DEFINITION 5.11. Let F be a 2SAT formula. A variable x is a *double companion* if and only if the degree of x in G_F is two.

For the following lemma, remember the definition of our new parameter t, the number of clause types.

LEMMA 5.12 (double companion reduction rule). *Let F be an arbitrary 2SAT formula. If x is a double companion, then we can transform F into a max-equivalent formula F' that contains the same variables as F except x, and possibly clauses of negative weight, in polynomial time. The formula F' does not have more clause types than F. Moreover, $G_{F'}$ is the graph obtained from G_F by contracting x.*

Proof. Let x be a double companion that occurs together with y and z. Let $F = F' + F''$, where F' consists of all the clauses that contain x and F'' holds all the other clauses. We define $a = OptVal(F'[y, z])$, $b = OptVal(F'[y, \bar{z}])$, $c = OptVal(F'[\bar{y}, z])$, and $d = OptVal(F'[\bar{y}, \bar{z}])$. Let

$$G = \{(a + b + c + d, \mathbf{T}), (-d, \{y, z\}), (-c, \{y, \bar{z}\}), (-b, \{\bar{y}, z\}), (-a, \{\bar{y}, \bar{z}\})\}.$$

We easily see $a = OptVal(G[y, z])$, $b = OptVal(G[y, \bar{z}])$, $c = OptVal(G[\bar{y}, z])$, and $d = OptVal(G[\bar{y}, \bar{z}])$. Therefore, $OptVal(F' + F'') = OptVal(G + F'')$. Moreover, x does obviously not occur in $G + F''$.

Note that the new clauses containing y and z imply the existence of an edge between the corresponding nodes in $G_{F'}$. Hence, $G_{F'}$ can be obtained from G_F by contracting x. ∎

We now have reduction rules for formulas in 2-CNF that enable us to eliminate all nodes with degree up to two in the corresponding connectivity graph. The following lemma shows how branching on a variable affects the connectivity graph.

LEMMA 5.13. *Let F be a formula and x a variable. Then $G_{F[x]} = G_{F[\overline{x}]} = G_F \setminus \{x\}$.*

Proof. Let F be a formula and x a variable. Setting x to true removes every clause containing x and shrinks each clause containing $\overline{x}$ to size one. Both operations result in the removal of all corresponding edges in G_F. Thus, $G_{F[x]} = G_F \setminus \{x\} = G_{F[\overline{x}]}$. ∎

DEFINITION 5.14. *Let F be a formula over variables $x_1, \ldots, x_n$ such that G_F is connected. If G_F is four-regular or its maximum degree does not equal four, the* preferable variable *is the first x_i of maximum degree. Otherwise, if G_F has maximum degree four as well as nodes of smaller degree, the* preferable variable *is the first x_i of degree four with at least one neighbor of smaller degree.*

From our results on confluence, we already know that a sequence of branching steps in Algorithm A is equivalent to branching on all involved variables and then applying the reduction rules as long as the formula does not decompose into several independent formulas. Whenever this is the case, each independent call of Algorithm A on F_i branches on a different set of variables. Observe that in this case,

simply branching on all variables could be too expensive—e.g., if each component is a clique of size five, we would have to branch on $m/5$ variables. Solving these formulas independently allows us to overcome this problem.

In order to ease the forthcoming proofs, we introduce the *splitting number* of a formula to represent the depth of this splitting process, i.e., how often a formula is separated on a single path in the recursion tree until all components are solved.

DEFINITION 5.15. *Let F be a formula. We define the* splitting number $s(F)$ *of a run of Algorithm A on F as*

(i) $s(F) := 0$ *if the algorithm returns in line* (2),
(ii) $s(F) := 1 + \max\{\, s(F_i) \mid i = 1, \ldots, l \,\}$ *if the algorithm returns in line* (4),
(iii) $s(F) := s(F[x])$ *if the algorithm returns in line* (7).

Clearly, $s(F[x]) = s(F[\overline{x}])$.

THEOREM 5.16. *Let F be a formula. Using only polynomial space, Algorithm A solves* MAX-2SAT *in time $O^*(2^{t/5.217})$ on F, where t is the number of clause types.*

Proof. Let $\Psi(F) = \Psi(G_F)$. Recall that branching on a variable x leads to the same connectivity graph in both branches, implying that the preferable variable is the same for both $F[x]$ and $F[\overline{x}]$. Therefore, Algorithm A always branches on the same variable set D until we end up with either $\{(k, T)\}$ or several independent subformulas F_i. In the latter case, new pairwise disjoint sets D of variables are used for each F_i. Note that the above variable sets D correspond to the node set D from Theorem 5.4 when applied to the respective connectivity graph, since Algorithm A selects the nodes in the same way as described in the proof of Theorem 5.4.

We prove a bound of $|F|2^{\Psi(F)/5.217}$ on the number of leaves in the recursion tree by induction over the splitting number $s(F)$, where $|F|$ denotes the number of variables in F.

If $s(F) = 0$, then $F = \{(k, T)\}$ and F never decomposes into several independent subformulas. Thus we obtain $R^*(G_F \setminus D) = \emptyset$. Consequently, $|D| \leq \Psi(G_F)/5.217 = \Psi(F)/5.217$, and therefore the number of leaves in the recursion tree is bounded by $2^{\Psi(F)/5.217}$.

If $s(F) > 0$, then F decomposes into several independent subformulas $F_1, \ldots, F_l$ after branching on all variables in D. By Theorem 5.4, the potential $\Psi(F_i)$ of each F_i is bounded by $\Psi(F) - |D| \cdot 5.217$. Obviously, we end up with $2^{|D|}$ different branches.

Using the induction hypothesis, we bound the number of recursive calls by

$$2^{|D|} \cdot \sum_{i=1}^{l} |F_i| \cdot 2^{(\Psi(F)-|D|\cdot 5.217)/5.217} = 2^{|D|} \cdot 2^{(\Psi(F)-|D|\cdot 5.217)/5.217} \cdot \sum_{i=1}^{l} |F_i| = |F| \cdot 2^{\Psi(F)/5.217}.$$

Since the number of leaves in the recursion tree is bounded by $|F| \cdot 2^{\Psi(F)/5.217}$, $\Psi(F)$ is a lower bound on the number of clause types, and each call takes only polynomial time, the running time of Algorithm A is $O^*(2^{t/5.217})$. ∎

Using the well-known reduction from MAX-CUT to MAX-2SAT which consists of two clauses for each edge but only one clause type, we obtain the following corollary.

COROLLARY 5.17. *Let G be a graph.* MAX-CUT *can be solved in at most $O^*(2^{m/5.217})$ steps on G using only polynomial space, where m denotes the number of edges in G.*

It is very simple to construct reduction rules for Max-Cut to deal with nodes of degree at most two. In doing so, the construction of a direct algorithm similar to Algorithm A for Max-Cut is straightforward.

Acknowledgments. We would like to thank Stefan Kratsch for pointing out a mistake in section 3 of the preliminary paper [10]. The mistake affects Lemma 6 of [10] and invalidates the proof of Theorem 2 of [10], which states a weaker bound of $m/5 + 2$ on the treewidth. We are also grateful to Gregory Sorkin for revealing a minor problem in our proof of the bound of $m/5.769 + O(\log n)$; the case where the graph decomposes into several parts has not been addressed in the original paper [10] (the completed proof can be found in section 4).

REFERENCES

[1] N. Bansal and V. Raman, *Upper bounds for MaxSat: Further improved*, in Proceedings of the 10th International Symposium on Algorithms and Computation (ISAAC), Lecture Notes in Comput. Sci. 1741, Springer-Verlag, New York, 1999, pp. 247–258.

[2] H. L. Bodlaender, *A tourist guide through treewidth*, Acta Cybernet., 11 (1993), pp. 1–21.

[3] H. L. Bodlaender, *A partial k-arboretum of graphs with bounded treewidth*, Theoret. Comput. Sci., 209 (1998), pp. 1–45.

[4] S. S. Fedin and A. S. Kulikov, *A $2^{|E|/4}$ algorithm for MAX-CUT*, J. Math. Sci., 126 (2005), pp. 1995–1999.

[5] F. V. Fomin and K. Høie, *Pathwidth of Cubic Graphs and Exact Algorithms*, Technical report 298, Department of Informatics, University of Bergen, Bergen, Norway, 2005.

[6] F. V. Fomin and K. Høie, *Pathwidth of cubic graphs and exact algorithms*, Inform. Process. Lett., 97 (2006), pp. 191–196.

[7] J. Gramm, E. A. Hirsch, R. Niedermeier, and P. Rossmanith, *New worst-case upper bounds for MAX-2-SAT with application to MAX-CUT*, Discrete Appl. Math., 130 (2003), pp. 139–155.

[8] T. Hofmeister, *An approximation algorithm for MAX-2-SAT with cardinality constraint*, in Proceedings of the 11th European Symposium on Algorithms (ESA), Lecture Notes in Comput. Sci. 2832, Springer-Verlag, New York, 2003, pp. 301–312.

[9] T. Kloks, *Treewidth*, Lecture Notes in Comput. Sci. 842, Springer-Verlag, New York, 1994.

[10] J. Kneis, D. Mölle, S. Richter, and P. Rossmanith, *Algorithms based on the treewidth of sparse graphs*, in Proceedings of the 31st International Workshop on Graph-Theoretic Concepts in Computer Science (WG), Lecture Notes in Comput. Sci. 3787, Springer-Verlag, New York, 2005, pp. 385–396.

[11] A. Kojevnikov and A. S. Kulikov, *A new approach to proving upper bounds for MAX-2-SAT,* in Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms (SODA), ACM, New York, SIAM, Philadelphia, 2006, pp. 11–17.

[12] A. S. Kulikov and K. Kutzkov, *New bounds for MAX-SAT by clause learning*, in Proceedings of the 2nd International Symposium on Computer Science in Russia, Ekaterinburg, Russia, 2007, pp. 194–204.

[13] D. Livnat M. Lewin and U. Zwick, *Improved rounding techniques for the MAX 2-SAT and MAX DI-CUT problems*, in Proceedings of the 9th Conference on Integer Programming and Combinatorial Optimization (IPCO), Lecture Notes in Comput. Sci. 2337, Springer-Verlag, New York, 2002, pp. 67–82.

[14] R. Niedermeier and P. Rossmanith, *New upper bounds for maximum satisfiability*, J. Algorithms, 36 (2000), pp. 63–88.

[15] D. Raible and H. Fernau, *A new upper bound for MAX-2-SAT: A graph-theoretic approach*, in Proceedings of the 33rd Conference on Mathematical Foundations of Computer Science (MFCS), E. Ochmanski and J. Tyszkiewicz, eds., Lecture Notes in Comput. Sci. 5162, Springer-Verlag, New York, 2008, pp. 551–562.

[16] A. D. Scott and G. B. Sorkin, *A faster exponential-time algorithm for Max 2-Sat, Max Cut, and Max k-Cut*, Technical report RC23456(W0412-001), IBM Research Report, 2004, available online at http://domino.research.ibm.com/library/cyberdig.nsf.

[17] A. D. Scott and G. B. Sorkin, *An LP-designed algorithm for constraint satisfaction*, in Proceedings of the 14th European Symposium on Algorithms (ESA), Lecture Notes in Comput. Sci. 4168, Springer-Verlag, New York, 2006, pp. 588–599.

[18] A. D. Scott and G. B. Sorkin, *Linear-programming design and analysis of fast algorithms for Max* 2-*SAT and Max* 2-*CSP*, Discrete Optim., 4 (2007), pp. 260–287.

[19] P. D. Seymour and R. Thomas, *Graph searching and a min-max theorem for tree-width*, J. Combin. Theory, 58 (1993), pp. 22–33.

[20] J. A. Telle and A. Proskurowski, *Algorithms for vertex partitioning problems on partial k-trees*, SIAM J. Discrete Math., 10 (1997), pp. 529–550.

[21] R. Williams, *A new algorithm for optimal constraint satisfaction and its implications*, in Proceedings of the 31st International Colloquium on Automata, Languages, and Programming (ICALP), Lecture Notes in Comput. Sci. 3142, Springer-Verlag, New York, 2004, pp. 1227–1237.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 428–446

MAXIMAL LABEL SEARCH ALGORITHMS TO COMPUTE PERFECT AND MINIMAL ELIMINATION ORDERINGS*

A. BERRY†, R. KRUEGER‡, AND G. SIMONET§

Abstract. Many graph search algorithms use a vertex labeling to compute an ordering of the vertices. We examine such algorithms which compute a peo (perfect elimination ordering) of a chordal graph and corresponding algorithms which compute an meo (minimal elimination ordering) of a non-chordal graph, an ordering used to compute a minimal triangulation of the input graph. We express all known peo-computing search algorithms as instances of a generic algorithm called MLS (maximal label search) and generalize Algorithm MLS into CompMLS, which can compute any peo. We then extend these algorithms to versions which compute an meo and likewise generalize all known meo-computing search algorithms. We show that not all minimal triangulations can be computed by such a graph search, and, more surprisingly, that all these search algorithms compute the same set of minimal triangulations, even though the computed meos are different. Finally, we present a complexity analysis of these algorithms.[1]

Key words. graph search, peo, meo, minimal triangulation, elimination scheme, maximal label search

AMS subject classification. 05C85

DOI. 10.1137/070684355

1. Introduction.

Graph searching plays a fundamental role in many algorithms, particularly using breadth-first or depth-first searches and their many variants. One important application is to compute special graph orderings related to the chordality of a graph. When the input graph is chordal, one wants to find an ordering of the vertices called a *peo* (perfect elimination ordering), which repeatedly selects a vertex whose neighborhood is a clique (called a *simplicial vertex*) and removes it from the graph. This is a certificate of chordality, as, given an ordering of the vertices, one can determine in linear time whether it is a peo of the graph.

When the input graph fails to be chordal, it is often interesting to embed it into a chordal graph by adding an inclusion-minimal set of edges, a process called *minimal triangulation*. One of the ways of accomplishing this is to use an ordering of the vertices called an *meo* (minimal elimination ordering) and use this to simulate a peo by repeatedly adding any edges whose absence would violate the simplicial condition.

Though some earlier work had been done on these problems (see [13, 12]), the seminal paper is that of Rose, Tarjan, and Lueker [14], which presented two very efficient algorithms to compute a peo or an meo. They introduced the concept of *lexicographic order* (which, roughly speaking, is a dictionary order) and used this for graph searches which at each step choose an unnumbered vertex of maximal label. With this technique, they introduced Algorithm LEX M, which for a non-chordal graph $G = (V, E)$ computes an meo in a very efficient $O(nm)$ time, (where $n = |V|$ and $m = |E|$) and then streamlined this for use on a chordal graph, introducing

*Received by the editors March 5, 2007; accepted for publication (in revised form) July 14, 2008; published electronically January 16, 2009.
http://www.siam.org/journals/sidma/23-1/68435.html
†LIMOS, Ensemble scientifique des Cézeaux, F-63177 Aubière, France (berry@isima.fr).
‡Department of Computer Science, University of Toronto, Toronto, Ontario, M5S 3G4 Canada (krueger@cs.toronto.edu).
§LIRMM, 161, Rue Ada, F-34392 Montpellier, France (simonet@lirmm.fr).
[1] An extended abstract of part of this paper was published in WG 2005 [4].

what is now called Algorithm LexBFS, a breadth-first search which runs in optimal $O(n + m)$ time and computes a peo if the input graph is chordal (see a survey on LexBFS in [7]).

Later work has been done on computing peos. Tarjan and Yannakakis [16] presented Algorithm MCS (maximal cardinality search), which is similar to LexBFS but uses a simplified labeling (a cardinality choice criterion is used instead of a lexicographic one). MCS also computes in linear time a peo if the input graph is chordal.

Shier [15] remarks that neither LexBFS nor MCS is capable of computing all peos. He proposes Algorithms MEC and MCC, that are generalizations of MNS and MCS, respectively, and can both compute any peo of a chordal graph.

Recently, Corneil and Krueger [8] introduced Algorithm LexDFS as a depth-first analogue to LexBFS. They also introduced Algorithm MNS (maximal neighborhood search), which chooses at each step a vertex whose set of numbered neighbors is inclusion-maximal. They gave characterizations of the orderings computed by these search algorithms and observed, from a result of Tarjan and Yannakakis [16] on the property characterizing MNS orderings, that every MNS ordering yields a peo if the input graph is chordal. They showed with these characterizations that any ordering computed by LexBFS, MCS, or LexDFS can also be computed by MNS.

Berry et al. [1] recently introduced Algorithm MCS-M, which computes an meo. MCS-M is extended from MCS in the same fashion LEX M can be extended from LexBFS. The sets of meos defined by LEX M and by MCS-M are different, but Villanger [19] recently showed that the same sets of minimal triangulations were obtained.

In this paper, we address natural questions which arise about peos and meos: how can the existing algorithms be generalized? Do these new algorithms compute all peos of a chordal graph? Can they all be extended to compute meos? What sets of minimal triangulations are obtained?

Algorithms LexBFS, MCS, LexDFS, and MNS clearly process in a similar way: they number the vertices of the input graph by repeatedly numbering an unnumbered vertex with maximal label and incrementing the labels of its neighbors. They only differ by their vertex labeling, i.e., the nature of labels and the way they are compared, initialized, and incremented. We show that they can be described as instances of a generic algorithm called MLS (maximal label search) having the vertex labeling as a parameter. We show that every instance of MLS computes a peo of a chordal graph but cannot compute every peo of every chordal graph. In order to obtain all possible peos, we extend MLS to CompMLS, which uses Shier's idea of working on the connected components of the subgraph induced by the unnumbered vertices. We show that every instance of generic CompMLS is capable of computing any peo of a chordal graph.

We then go on to examine the issues pertaining to meos and minimal triangulations. We show that MNS, MLS, and CompMLS can all be extended to compute an meo, in the same way that LEX M is extended from LexBFS. We show the very strong result that all the sets of minimal triangulations computed are the same, independent of the meo-computing algorithm which is used, and that not all minimal triangulations can be computed by this new family of algorithms.

The paper is organized as follows: in section 2 we give some definitions and notations, in section 3 we discuss peos, in section 4 we discuss meos, and in section 5 we present a complexity analysis of the algorithms defined in the paper.

2. Preliminaries. All graphs in this work are undirected and finite. A graph is denoted $G = (V, E)$, with $n = |V|$ and $m = |E|$. The *neighborhood* of a vertex x in G is denoted $N_G(x)$, or simply $N(x)$ if the meaning is clear. An ordering on V is a one-to-one mapping from $\{1, 2, \ldots, n\}$ to V. In every figure in this paper showing

an ordering α on V, α is defined by giving on the figure the number $\alpha^{-1}(x)$ for every vertex x. $\mathbb{Z}^+$ denotes the set of positive integers $\{1, 2, 3, \dots\}$, and for any positive integer i, $\mathbb{Z}_i^+$ denotes the set of positive integers strictly larger than i.

A *chordal* (or *triangulated*) graph is a graph with no chordless cycle of length greater or equal to 4. To recognize chordal graphs efficiently, Fulkerson and Gross [11] used a greedy elimination structure on simplicial vertices: "A graph is chordal iff one can repeatedly find a simplicial vertex and delete it from the graph, until no vertex is left" (a vertex is simplicial if its neighborhood is a clique). This defines an ordering on the vertices which is called a *peo* of the graph.

When a graph G fails to be chordal, any ordering α on the vertices can be used to embed G into a chordal graph (called a *triangulation* of G) by repeatedly choosing the next vertex x, adding any edges necessary to make it simplicial, and removing x. If F is the set of added edges, the graph obtained is chordal and is denoted $H = (V, E + F) = G_\alpha^+$.

If $H = (V, E + F)$ is a triangulation of $G = (V, E)$ and if for every proper subset $F' \subset F$, graph $(V, E + F')$ fails to be chordal, H is called a *minimal triangulation* of G. If, moreover, α is an ordering such that $H = G_\alpha^+$, α is called a *meo* of G.

In [14], two very important characterizations are given.

Path Lemma. For any graph $G = (V, E)$, any ordering α on V, and any x, y in V such that $\alpha^{-1}(y) < \alpha^{-1}(x)$, xy is an edge of G_α^+ iff there is a path μ in G from x to y such that $\forall t \in \mu \setminus \{x, y\},\ \alpha^{-1}(t) < \alpha^{-1}(y)$.

Unique Chord Property.

For any graph $G = (V, E)$ and any triangulation $H = (V, E + F)$ of G, H is a minimal triangulation of G iff each edge in F is the unique chord of a 4-cycle of H.

3. Computing peos. Every one of Algorithms LexBFS, MCS, LexDFS, and MNS works in the following fashion: Start with a graph where all vertices are unnumbered and have the same label. Repeatedly choose an unnumbered vertex x whose label is maximal (with respect to a given partial order on labels), give x the following number i (in increasing or decreasing order according to the algorithm), and increment the label of each as yet unnumbered neighbor of x into a new value depending on its current value and i. Algorithms LexBFS and MCS, as defined in [14] and [16], number vertices from n down to 1, whereas LexDFS and MNS, as defined in [8], number vertices from 1 to n, so that they actually compute the reverse of a peo of every chordal graph. In this paper, our algorithms compute peos and meos directly, and thus number vertices from n down to 1; thus, vertex number 1 of a peo-computing algorithm will be a simplicial vertex of the graph.

In order to define a generic peo-computing algorithm, we first define a *labeling structure*.

DEFINITION 3.1. *A* labeling structure *is a structure* $(L, \preceq, l_0, Inc)$*, where*

- L *is a set (the set of labels);*
- $\preceq$ *is a partial order on* L *(which may be total or not, with* $\prec$ *denoting the corresponding strict order), which will be used to choose a vertex of maximal label;*
- l_0 *is an element of* L*, which will be used to initialize the labels;*
- Inc *is a mapping from* $L \times \mathbb{Z}^+$ *to* L*, which will be used to increment a label, and such that the following condition IC (inclusion condition) holds:*
  IC : for any subsets I *and* I' *of* $\mathbb{Z}_2^+$*, if* $I \subset I'$*, then* $lab_S(I) \prec lab_S(I')$*, where* $lab_S(I) = Inc(\dots(Inc(Inc(l_0, i_1), i_2), \dots), i_k)$*, where* $I = \{i_1, i_2, \dots, i_k\}$*, with* $i_1 > i_2 > \cdots > i_k$*.*

Algorithm MLS (maximal label Search)
input : A graph $G = (V, E)$ and a labeling structure $(L, \preceq, l_0, Inc)$.
output: An ordering α on V.
Initialize all labels as l_0; $G' \leftarrow G$;
for $i = n$ **downto** 1 **do**
  Choose a vertex x of G' of maximal label;
  $\alpha(i) \leftarrow x$;
  foreach y *in* $N_{G'}(x)$ **do**
    $label(y) \leftarrow Inc(label(y), i)$;
  Remove x from G';

FIG. 1. *Algorithm MLS.*

It will sometimes be useful to use the number n of vertices of the graph to be labeled in the definition of $Inc(l, i)$. It will be the case, for instance, for the labeling structure S_3 associated with Algorithm LexDFS. In that case, Inc can be seen as a family of mappings Inc_n from $L \times \mathbb{Z}^+$ to L for each positive integer n.

The corresponding algorithm, which we introduce as MLS, is given by Figure 1. MLS iteratively selects a vertex to add to the ordering and increments the labels of its unselected neighbors. We will refer to the iteration of the loop that defines $\alpha(i)$ as Step i of the algorithm.

LexBFS, MCS, LexDFS, and MNS are all special cases of MLS, with the following labeling structures $(L, \preceq, l_0, Inc)$; in each case, we also give the value of $lab_S(I)$ for any subset I of $\mathbb{Z}^+$.

LexBFS (Structure S_1): L is the set of lists of elements of $\mathbb{Z}^+$, $\preceq$ is the lexicographic order (a total order), l_0 is the empty list, $Inc(l, i)$ is obtained from l by adding i to the end of the list, $lab_{S_1}(I)$ is the string of the integers in I in decreasing order.

MCS (Structure S_2): $L = \mathbb{Z}^+ \cup \{0\}$, $\preceq$ is $\leq$ (a total order), $l_0 = 0$, $Inc(l, i) = l+1$, $lab_{S_2}(I) = |I|$.

LexDFS (Structure S_3): L is the set of lists of elements of $\mathbb{Z}^+$, $\preceq$ is the lexicographic order (a total order), l_0 is the empty list, $Inc(l, i)$ is obtained from l by adding $n+1-i$ to the beginning of the list, $lab_{S_3}(I)$ is the string of the complements to $n+1$ of the integers in I in decreasing order.

MNS (Structure S_4): L is the power set of $\mathbb{Z}^+$, $\preceq$ is $\subseteq$ (not a total order), $l_0 = \emptyset$, $Inc(l, i) = l \cup \{i\}$, $lab_{S_4}(I) = I$.

In our proofs, we will use the following notations.

NOTATIONS 3.2. *For any graph $G = (V, E)$, any execution of our algorithms on G computing some ordering α on V, and any integer i between* 1 *and* n,

- V_i is the set of still unnumbered vertices at the beginning of Step i, i.e., the set $\{\alpha(j),\ 1 \leq j \leq i\}$;

- G'_i is graph G' at the beginning of Step i, i.e., the subgraph of G induced by V_i, and, for each $y \in V_i$,

- $label_i(y)$ is the value of $label(y)$ at the beginning of Step i and

- $Num_i(y) = \{j \in \{i+1, i+2, \ldots, n\} \mid label(y)$ has been incremented at Step $j\}$.

The following Lemma is clear from Algorithm MLS.

LEMMA 3.3. *For any graph $G = (V, E)$, any labeling structure S, any execution of MLS on G and S computing some ordering α on V, any integer i between* 1 *and* n, *and any $y \in V_i$, $label_i(y) = lab_S(Num_i(y))$ and $Num_i(y) = Num^{\alpha}_{G,i}(y)$, where $Num^{\alpha}_{G,i}(y)$ denotes the set of integers $j > i$ such that $\alpha(j)$ is adjacent to y in G.*

Thus, the label of y at the beginning of Step i is equal to $lab_S(Num^{\alpha}_{G,i}(y))$, where $Num^{\alpha}_{G,i}(y)$ is defined from the ordering α computed so far on numbered vertices, independently from the labeling structure involved. This property will allow us to characterize the orderings computed by MLS and to compare the sets of orderings computed with different labeling structures (Characterization 3.4 and Lemma 3.5).

We can view MLS as a generic algorithm with parameter S. For every labeling structure S, we denote by S-MLS the instance of generic Algorithm MLS using this particular labeling structure S and by "S-MLS ordering of a graph G" any ordering that can be computed by S-MLS on input graph G. Thus, LexBFS is S_1-MLS, MCS is S_2-MLS, LexDFS is S_3-MLS, and MNS is S_4-MLS.

The set of S-MLS orderings of a given graph depends on S. An MLS ordering of a graph G is an ordering that can be computed by MLS on G, i.e., by S-MLS for some labeling structure S. Thus, the set of MLS orderings of G is the union of the sets of S-MLS orderings of G for all labeling structures S.

The following theorem shows that MNS can compute every S-MLS ordering of a given graph for every labeling structure S. This theorem can be proved using the MNS characterization presented in [8]. We will prove it by using the following more general results.

CHARACTERIZATION 3.4. *For any graph G, any labeling structure S, and any ordering α of V, α is an S-MLS ordering of G iff for any integers i, j such that $1 \leq j < i \leq n$, $lab_S(Num^{\alpha}_{G,i}(\alpha(i))) \not\prec lab_S(Num^{\alpha}_{G,i}(\alpha(j)))$.*

Proof. α is an S-MLS ordering of G iff for any integer i between 1 and n, the label of $\alpha(i)$ at the beginning of Step i is maximal among the labels of vertices $\alpha(j)$, $i \leq j \leq n$. We conclude with Lemma 3.3. ∎

LEMMA 3.5. *Let S and S' be labeling structures with partial orders $\preceq_S$ and $\preceq_{S'}$, respectively, such that for any subsets I and I' of $\mathbb{Z}_2^+$, if $lab_{S'}(I) \prec_{S'} lab_{S'}(I')$, then $lab_S(I) \prec_S lab_S(I')$.*

Then every S-MLS ordering of G is also an S'-MLS ordering of G for every graph G.

Proof. Let G be a graph and α be an S-MLS ordering of G. By Characterization 3.4, for any integers i, j such that $1 \leq i < j \leq n$, $lab_S(Num^{\alpha}_{G,i}(\alpha(i))) \not\prec_S lab_S(Num^{\alpha}_{G,i}(\alpha(j)))$, where $Num^{\alpha}_{G,i}(\alpha(i))$ and $Num^{\alpha}_{G,i}(\alpha(j))$ are subsets of $\mathbb{Z}_2^+$ since $i > 1$, so $lab_{S'}(Num^{\alpha}_{G,i}(\alpha(i))) \not\prec_{S'} lab_{S'}(Num^{\alpha}_{G,i}(\alpha(j)))$. By Characterization 3.4 again, α is an S'-MLS ordering of G. ∎

THEOREM 3.6. *For any graph $G = (V, E)$ and any labeling structure S, any S-MLS ordering of G is an MNS ordering of G.*

Proof. This follows immediately from Lemma 3.5 and condition IC, since MNS $= S_4$-MLS, with $lab_{S_4}(I) = I$ and $\prec_{S_4} = \subset$. ∎

A corollary of Theorem 3.6 is that any instance of MLS computes a peo of a chordal graph, since this is true for MNS [8].

Another consequence is that any LexBFS, MCS, or LexDFS ordering of a graph is also an MNS ordering, which already follows from the characterizations given in [8]. However, for arbitrary labeling structures S and S', an ordering computed by S-MLS need not be computable by S'-MLS. For instance, Figure 2(a) shows a LexBFS ordering which is not an MCS ordering, while Figure 2(b) shows an MCS ordering which is not a LexBFS ordering. There also exist graphs with MNS orderings that are neither LexBFS nor MCS orderings.

As the set of MLS orderings of a graph G is the union of the sets of S-MLS orderings of G for all labeling structures S and as MNS is equal to S_4-MLS, it follows

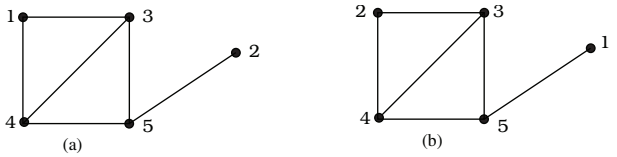


FIG. 2. *A chordal graph with different* (a) *LexBFS and* (b) *MCS orderings.*

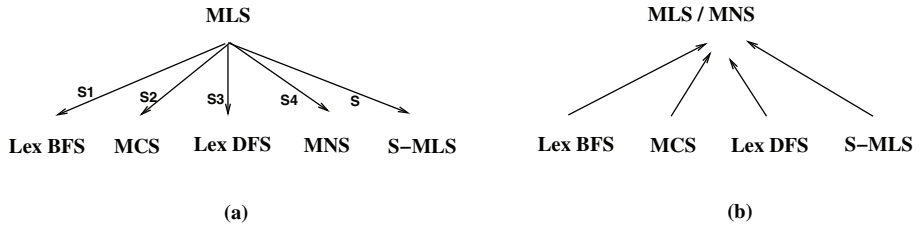


FIG. 3. (a) *Instances of MLS and* (b) *Inclusion order on the sets of computable orderings.*

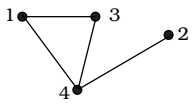


FIG. 4. α *is a CompMNS ordering of* G *but not an MNS one.*

from Theorem 3.6 that every graph has the same MLS and MNS orderings. However, be careful that MLS and MNS are different algorithms, since MLS has a graph and a labeling structure as input, whereas MNS only has a graph (or, if MLS is seen as a generic algorithm with a labeling structure as a parameter, MNS is only an instance of MLS). Figure 3 gives two different relations on peo-computing algorithms. Figure 3(a) shows some instances of generic Algorithm MLS, each arrow from MLS to one of its instances being labeled with the corresponding value of parameter S. Figure 3(b) shows the inclusion order on the sets of orderings computable by these algorithms on a given graph. In this figure, S is an arbitrary labeling structure.

It is interesting to remark that even though MLS, or equivalently MNS, is more general (in the sense that it can compute more peos) than LexBFS and MCS, it still is not powerful enough to compute every possible peo of a given chordal graph. This is shown by the simple counterexample in Figure 4: no MLS execution on this graph will find the ordering indicated, although it is clearly a peo.

In order to make it possible to find *any* peo, we further generalize MLS using Shier's idea [15] of using the *connected components* of the subgraph G' induced by the unnumbered vertices. We thus introduce Algorithm CompMLS, defined from Algorithm MLS, by replacing the following:
"Choose a vertex x of G' of maximal label;" with
"Choose a connected component C of G';
"choose a vertex x of C of maximal label in C;".

This generalizes the entire family of peo-computing algorithms discussed in this paper: for any X in {LexBFS, MCS, LexDFS, MNS, MLS}, Algorithm CompX is a generalization of X, and we will show that it computes a peo if the graph is chordal.

Algorithms MEC and MCC defined by Shier [15] are instances of generic Algorithm CompMLS: MEC is CompMNS, i.e., S_4- CompMLS, and MCC is CompMCS, i.e., S_2-CompMLS. Algorithm CompMNS can compute the peo of Figure 4. In fact, Shier proved in [15] that CompMNS and even CompMCS compute all peos of a chordal graph. We show that this holds for every instance of Algorithm CompMLS, using some results from section 4.

THEOREM 3.7. *For any chordal graph G and any labeling structure S, the S-CompMLS orderings of G are exactly its peos.*

Proof. Let G be a chordal graph and S be a labeling structure. By [15], the CompMNS orderings of G are exactly its peos, and by Theorem 4.15 from section 4, G has the same S-CompMLSM and CompMNSM orderings, which are also its S-CompMLS and CompMNS orderings, since G is chordal (by the extension of Property 4.7 from section 4 to CompMLS and CompMLSM). ∎

We consider here that a labeling structure is defined without the condition IC, and we discuss the choice of the condition IC in view of obtaining an algorithm computing peos of chordal graphs. By Theorem 3.6, IC is a sufficient condition on a labeling structure S for S-MLS to compute only peos of every chordal graph. It turns out that it is also a necessary one, so that IC is exactly the condition required on a labeling structure for MLS to compute only peos of every chordal graph.

THEOREM 3.8. *The condition IC imposed on a labeling structure S is a necessary and sufficient condition for S-MLS to compute only peos of every chordal graph.*

Proof. IC is a sufficient condition, since by Theorem 3.6, S-MLS computes only MNS orderings, and therefore peos, of every chordal graph.

Conversely, suppose there are some subsets I and I' of $\mathbb{Z}_2^+$ such that $I \subset I'$ and $lab_S(I) \not\prec lab_S(I')$ and let us show that there is a chordal graph G and an S-MLS ordering of G that is not a peo of G. Let $q = \max(I')$. We choose two subset I and I' of $\mathbb{Z}_2^+$ such that $I \subset I'$, $lab_S(I) \not\prec lab_S(I')$, $\max(I') = q$ and $\min(I')$ is the largest possible with these conditions. Let $p = \min(I')$. $p > 2$ since I' is a subset of of $\mathbb{Z}_2^+$. Let $G = (V, E)$, with $V = \{z_1, z_2, \ldots, z_q\}$ and $E = \{z_i z_j, p \leq i < j \leq q\} \cup \{z_i z_{p-1}, i \in I'\} \cup \{z_i z_{p-2}, i \in I\} \cup \{z_{p-1} z_{p-2}\}$. G is chordal since $(z_1, z_2, \ldots, z_q)$ is a peo of G. By the choice of I and I', there is an execution of S-MLS on G choosing $z_q, z_{q-1}, \ldots, z_p$ first and then choosing z_{p-2} before z_{p-1}. The resulting S-MLS ordering of G is not a peo of G because the set of neighbors of z_{p-1} with higher numbers than z_{p-1} in this ordering is not a clique, since z_{p-2} is not adjacent to the vertices of the form z_i, $i \in I' \setminus I$. ∎

The condition imposed on a labeling structure was defined in a different way in [4]. Instead of satisfying IC, the mapping Inc had to satisfy the following condition: For any integer n in $\mathbb{Z}^+$, any integer i between 1 and n and any labels l and l' in L_i^n, the following properties hold:

(ls1) $l \prec Inc(l, i)$;

(ls2) if $l \prec l'$, then $Inc(l, i) \prec Inc(l', i)$, where L_i^n is the subset of L defined by induction on i by

$L_n^n = \{l_0\}$ and $L_{i-1}^n = L_i^n \cup \{l = Inc(l', i) \mid l' \in L_i^n\}$ for any i from n down to 2.

It is easy to show that this condition implies IC, but the converse is not true. For instance, let S be the labeling structure obtained from S_4 (the structure used for MNS) by replacing the inclusion order $\preceq_{S_4}$ by the partial order $\preceq$ on L defined by the following: For any $l, l' \in L$, $l \preceq l'$ iff ($l \subseteq l'$ or ($l = \{4\}$ and $5 \in l'$)) (checking that $\preceq$ is a partial order on L is left to the reader). IC holds since the inclusion order is a suborder of $\preceq$, but not (ls2) since $\{4\} \prec \{5\}$ but $Inc(\{4\}, 3) = \{4, 3\} \not\prec \{5, 3\} = Inc(\{5\}, 3)$. Thus, IC is more appropriate than the conjunction of (ls1) and (ls2) in the context

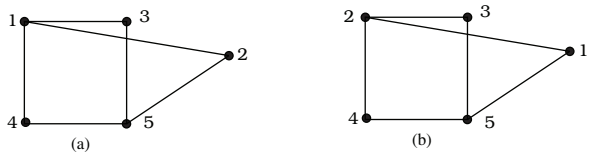


FIG. 5. *A non-chordal graph with different* (a) *CompLexBFS and* (b) *CompMCS orderings.*

of peo-computing algorithms, though a labeling structure satisfying IC necessarily satisfies (ls1) and often satisfies (ls2) in practice.

Let us conclude this section by some remarks on running the MLS family of algorithms on *non-chordal* graphs. LexBFS has been used on asteroidal triple-free graphs [9] and has been shown to have very interesting invariants even on an arbitrary graph (see [2, 3]). Likewise, MCS has also been used on various graph classes (see [10, 6]).

Unlike a chordal graph, a non-chordal graph does not necessarily have the same CompLexBFS, CompMCS, CompLexDFS, and CompMNS orderings. Figure 5(a) shows a CompLexBFS ordering which is not a CompMCS one, while Figure 5(b) shows a CompMCS ordering which is not a CompLexBFS one.

4. Computing meos. We will now introduce the extensions of Algorithms MNS, MLS, and CompMLS into their meo-computing counterparts.

To extend LexBFS into LEX M, at each step choosing a vertex x of maximum label $label(x)$, an edge is added between x and any unnumbered vertex y whenever there is a path from x to y in the subgraph induced by the unnumbered vertices such that all internal vertices on the path have a label strictly smaller than the label of y. This approach has been used recently in [1] to extend MCS into meo-computing Algorithm MCS-M; here, we extend MLS into MLSM, as given by Figure 6. Thus, LEX M is S_1-MLSM, MCS-M is S_2-MLSM, LexDFS-M is defined as S_3-MLSM, and MNSM is defined as S_4-MLSM. We will see that Algorithm MNSM is, in fact, as general as MLSM: every MLSM ordering of a graph is an MNSM ordering.

For any labeling structure S, we call S-MLSM the instance of Algorithm MLSM using S, and S-MLSM ordering an ordering computed by S-MLSM.

Clearly, the relation between $label_i(y)$ and $Num_i(y)$ in an execution of MLS still holds in an execution of MLSM.

LEMMA 4.1. *For any graph $G = (V, E)$, any labeling structure S, any execution of MLSM on G and S, any integer i between* 1 *and n, and any $y \in V_i$,*

$$label_i(y) = lab_S(Num_i(y)).$$

We will show that as for MLS, $Num_i(y)$ can be defined from the ordering α computed so far on numbered vertices, independently from the labeling structure involved, with similar consequences in terms of characterizing MLSM orderings and comparing the sets of orderings computed by MLSM with different labeling structures.

4.1. The MLSM family of algorithms.

THEOREM 4.2. *For any execution of MLSM, $H = G^+_\alpha$ and α is a meo of G.*

To prove this, we will need several technical lemmas. Lemma 4.3 is clear from algorithm MLSM, Lemma 4.4 immediately follows from Lemma 4.1 and condition IC. The proof of Lemma 4.5 is long and technical and so, for reasons of readability, is given in the Appendix.

Algorithm MLSM (maximal label search for meo).

input : A graph $G = (V, E)$ and a labeling structure $(L, \preceq, l_0, Inc)$.
output: An meo α on V and a minimal triangulation $H = G^+_\alpha$ of G.

Initialize all labels as l_0; $E' \leftarrow \emptyset$; $G' \leftarrow G$;
for $i = n$ **downto** 1 **do**
  Choose a vertex x of G' of maximal label;
  $\alpha(i) \leftarrow x$;
  foreach *vertex* y *of* G' *different from* x **do**
    if *there is a path from* x *to* y *in* G' *such that every internal vertex on the path has a label strictly smaller than* $label(y)$, **then**
      $E' \leftarrow E' \cup \{xy\}$;
  foreach y *in* V *such that* $xy \in E'$ **do**
    $label(y) \leftarrow Inc(label(y), i)$;
  Remove x from G';
$H \leftarrow (V, E')$;

FIG. 6. *Algorithm MLSM.*

LEMMA 4.3. *For any graph $G = (V, E)$, any execution of MLSM on G computing ordering α and graph H, any integers i, j such that $1 \le i < j \le n$, and any y in V_i, the following propositions are equivalent:*

1. $j \in Num_i(y)$,
2. $\alpha(j)y$ *is an edge of* H,
3. *There is a path μ in G'_j from $\alpha(j)$ to y such that $\forall t \in \mu \setminus \{\alpha(j), y\}$, $label_j(t) \prec label_j(y)$.*

LEMMA 4.4. *For any graph $G = (V, E)$, any execution of MLS or MLSM on G, any integer i between 1 and n, and any x, y in V_i,*

(i) *if $Num_i(x) = Num_i(y)$, then $label_i(x) = label_i(y)$, and*
(ii) *if $Num_i(x) \subset Num_i(y)$, then $label_i(x) \prec label_i(y)$.*

LEMMA 4.5. *For any graph G, any execution of MLSM on G computing ordering α, any integer i between 1 and n, and any path μ in G'_i ending in some vertex y,*

(a) $\forall t \in \mu \setminus \{y\}$, $label_i(t) \prec label_i(y)$ *iff* $\forall t \in \mu \setminus \{y\}$, $Num_i(t) \subset Num_i(y)$;
(b) *if* $\forall t \in \mu \setminus \{y\}$, $label_i(t) \prec label_i(y)$, *then* $\forall t \in \mu \setminus \{y\}$, $\alpha^{-1}(t) < \alpha^{-1}(y)$;
(c) *if* $\forall t \in \mu \setminus \{y\}$, $\alpha^{-1}(t) < \alpha^{-1}(y)$, *then* $\forall t \in \mu \setminus \{y\}$, $Num_i(t) \subseteq Num_i(y)$.

Proof of Theorem 4.2. We first show that for any execution of MLSM, $H = G^+_\alpha$. Let $x, y \in V$ such that $\alpha^{-1}(y) < \alpha^{-1}(x) = i$. Let us show that xy is an edge of H iff it is an edge of G^+_α.

If xy is an edge of H, then, by Lemma 4.3, there is a path μ in G'_i from x to y such that $\forall t \in \mu \setminus \{x, y\}, label_i(t) \prec label_i(y)$. By Lemma 4.5 (b), $\forall t \in \mu \setminus \{x, y\}$, $\alpha^{-1}(t) < \alpha^{-1}(y)$, and, by the path lemma, xy is an edge of G^+_α.

Conversely, let xy be an edge of G^+_α. Let us show that xy is an edge of H. By the path lemma, there is a path μ in G from x to y such that $\forall t \in \mu \setminus \{x, y\}$, $\alpha^{-1}(t) < \alpha^{-1}(y) < i$, so $\mu \setminus \{x\} \subseteq V_{i-1}$. By Lemma 4.5(c), $\forall t \in \mu \setminus \{x, y\}$, $Num_{i-1}(t) \subseteq Num_{i-1}(y)$. Let t_1 be the neighbor of x in μ. xt_1 is an edge of H, so, by Lemma 4.3, $i \in Num_{i-1}(t_1)$; hence, $i \in Num_{i-1}(y)$, and, by Lemma 4.3, xy is an edge of H.

We now show that G^+_α is a minimal triangulation of G. Let $H = G^+_\alpha = (V, E+F)$. As G^+_α is a triangulation of G, by the unique chord property, it is sufficient to show that each edge in F is the unique chord of a cycle in H of length 4. Let xy be an edge in F, with $\alpha^{-1}(y) < \alpha^{-1}(x) = i$. xy is an edge of H, so, by Lemma 4.3, there

is a path μ in G'_i from x to y such that $\forall t \in \mu \setminus \{x,y\}$, $label_i(t) \prec label_i(y)$, and also $\alpha^{-1}(t) < \alpha^{-1}(y)$ by Lemma 4.5(b). $\mu \setminus \{x,y\} \neq \emptyset$, since xy is not an edge in G. Let t_1 be the vertex in $\mu \setminus \{x,y\}$ such that $\alpha^{-1}(t_1)$ is maximum. By the path lemma, xt_1 and t_1y are edges of G^+_α and, therefore, of H. As $label_i(t_1) \prec label_i(y)$, by Lemma 4.4, $Num_i(y) \not\subseteq Num_i(t_1)$. Let $j \in Num_i(y) \setminus Num_i(t_1)$ and $z = \alpha(j)$. $j > i$, and, by Lemma 4.3, yz is an edge of H (and, therefore, of G^+_α), but t_1z is not. Since yx and yz are edges of G^+_α with $\alpha^{-1}(y) < \alpha^{-1}(x) = i < j = \alpha^{-1}(z)$, by the definition of G^+_α, xz is an edge of G^+_α, and, therefore, of H. Hence, xy is the unique chord of cycle (x, t_1, y, z, x) in H of length 4. □

Thus, MLSM (and also LEX M, MCS-M, LexDFS-M, and MNSM) computes an meo and a minimal triangulation of the input graph. An execution of MLSM has the same behavior (same labeling and numbering) on the input graph G as an execution of MLS on the output minimal triangulation G^+_α, breaking ties in the same way. If, moreover, G is chordal, then G is equal to its minimal triangulation G^+_α, so that MLS and MLSM have the same behavior on G. We immediately obtain the following two properties.

PROPERTY 4.6. *For any graph G and any labeling structure S, any S-MLSM ordering α of G is an S-MLS ordering of G^+_α.*

PROPERTY 4.7. *For any chordal graph G and any labeling structure S, G has the same S-MLS and S-MLSM orderings.*

It follows from Property 4.7 and Theorems 3.8 and 4.2 that IC is exactly the condition required on a labeling structure S for S-MLSM to compute only meos of every graph.

COROLLARY 4.8. *Condition IC imposed on a labeling structure S is a necessary and sufficient condition for S-MLSM to compute only meos of every graph.*

Proof. IC is a sufficient condition by Theorem 4.2.

Conversely, if S-MLSM computes only meos of every graph, then it computes only peos of every chordal graph and so does S-MLS by Property 4.7. It follows by Theorem 3.8 that S satisfies IC. □

Let us remark that, for two given structures S and S', the sets of orderings computed by S-MLSM and S'-MLSM may be different, as is the case for S-MLS and S'-MLS. LEX M and MCS-M, for example, compute different orderings, as shown in Figure 2 (since MLS and MLSM compute the same orderings on a chordal graph). In the same way that MNS is as general as MLS, it turns out that MNSM is as general as MLSM, thus every graph has the same MLSM and MNSM orderings. The proof goes as for MLS and MNS, since by Lemma 4.5(a), $Num_i(y)$ can be defined from the ordering α computed on numbered vertices, independently from the labeling structure used.

LEMMA 4.9. *For any graph $G = (V, E)$, any execution of MLSM on G computing some ordering α on V, any integer i between 1 and n, and any $y \in V_i$, $Num_i(y)) = NumM^\alpha_{G,i}(y))$, where $NumM^\alpha_{G,i}$ is defined on V_i by induction on i from n down to 1 by the following:*

$NumM^\alpha_{G,n}(y) = \emptyset$, and for any i from n down to 2, $NumM^\alpha_{G,i-1}(y) = NumM^\alpha_{G,i}(y) \cup \{i\}$ if there is a path μ in G'_i from $\alpha(i)$ to y such that $\forall t \in \mu \setminus \{\alpha(i), y\}$, $NumM^\alpha_{G,i}(t) \subset NumM^\alpha_{G,i}(y)$, otherwise $NumM^\alpha_{G,i-1}(y) = NumM^\alpha_{G,i}(y)$.

CHARACTERIZATION 4.10. *For any graph G, any labeling structure S, and any ordering α of V, α is an S-MLSM ordering of G iff for any integers i, j such that $1 \leq j < i \leq n, lab_S(NumM^\alpha_{G,i}(\alpha(i))) \not\prec lab_S(NumM^\alpha_{G,i}(\alpha(j)))$.*

LEMMA 4.11. *Let S and S' be labeling structures with partial orders $\preceq_S$ and $\preceq_{S'}$, respectively, such that for any subsets I and I' of $\mathbb{Z}^+_2$, if $lab_{S'}(I) \prec_{S'} lab_{S'}(I')$, then $lab_S(I) \prec_S lab_S(I')$.*

Then every S-MLSM ordering of G is also an S′-MLSM ordering of G for every graph G.

THEOREM 4.12. *For any graph $G = (V, E)$ and any labeling structure S, any S-MLSM ordering of G is an MNSM ordering of G.*

4.2. The CompMLSM family of algorithms. We define CompMLSM from MLSM in the same way we defined CompMLS from MLS. Properties extend readily from an MLSM algorithm to its CompMLSM version: at Step i, our proofs only compare the label of $\alpha(i)$ to labels of vertices along paths in the graph G'_i, so $\alpha(i)$ needs only be maximal within the connected component of G'_i containing it.

We thus have similar results.

THEOREM 4.13. *For any input graph G and any X in {LEX M, MCS-M, LexDFS-M, MNSM, MLSM}, CompX computes an meo α of G and the associated minimal triangulation G^+_α of G.*

We also easily extend results such as Properties 4.6 and 4.7 and Characterization 4.10.

An important difference between MLSM and CompMLSM is that the set of orderings CompMLSM can find is independent of the labeling structure used and is a superset of the set of orderings obtainable by any algorithm of the MLSM family.

LEMMA 4.14. *For any execution of MLSM or CompMLSM on a graph G, any integer i between 1 and n, and any vertex y of the connected component of G'_i containing $\alpha(i)$, $Num_i(y) \subseteq Num_i(\alpha(i))$.*

Proof. Let μ be a path in G'_i between $\alpha(i)$ and y. $\forall t \in \mu \setminus \{\alpha(i)\},\ \alpha^{-1}(t) < i$, so, by Lemma 4.5(c), $Num_i(y) \subseteq Num_i(\alpha(i))$. ∎

THEOREM 4.15. *Any graph has the same S-CompMLSM orderings for all labeling structures S.*

Proof. Let G be a graph and S, S' be labeling structures. Let α be an S-CompMLSM ordering of G, let us show that α is an S′-CompMLSM ordering of G. By the extension of Characterization 4.10 to CompMLSM, it is sufficient to show that for any integers i, j such that $1 \le j < i \le n$ and $\alpha(i)$ and $\alpha(j)$ are in the same connected component of the subgraph of G induced by $\{\alpha(k),\ 1 \le k < i\}$, $lab_{S'}(NumM^\alpha_{G,i}(\alpha(i))) \not\prec lab_{S'}(NumM^\alpha_{G,i}(\alpha(j)))$. Let i, j be such integers. By Lemma 4.14, $Num_i(\alpha(j)) \subseteq Num_i(\alpha(i))$ in an execution of S-CompMLSM computing α, so, by Lemma 4.9, $NumM^\alpha_{G,i}(\alpha(j)) \subseteq NumM^\alpha_{G,i}(\alpha(i))$. It follows by condition IC that $lab_{S'}(NumM^\alpha_{G,i}(\alpha(j))) \preceq lab_{S'}(NumM^\alpha_{G,i}(\alpha(i)))$, and therefore

$$lab_{S'}(NumM^\alpha_{G,i}(\alpha(i))) \not\prec lab_{S'}(NumM^\alpha_{G,i}(\alpha(j))).$$ ∎

Every chordal graph G has the same S-CompMLSM and S-CompMLS orderings, which are exactly its peos (Theorem 3.7, whose proof uses Theorem 4.15).

Computing all peos does not extend to meos for the MLSM family of algorithms: Figure 7 shows an meo which CompMLSM is not capable of computing.

This raises the question of which minimal triangulations can be obtained by various algorithms of this family. Villanger in [19] proved the surprising result that the sets of minimal triangulations obtainable by LEX M and MCS-M are the same. Upon investigation, it turns out that, given one of these algorithms, using its Comp version does not enlarge the set of computable triangulations, although the set of computable meos may be larger.

THEOREM 4.16. *For any graph G and any given labeling structure S, G has the same sets of S-MLSM and of S-CompMLSM minimal triangulations.*

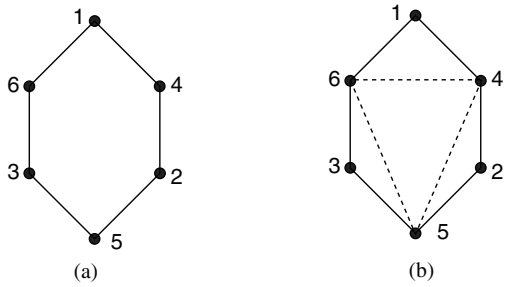


FIG. 7. (a) *Graph G and an meo α of G.* (b) *The corresponding minimal triangulation G^+_α of G. No version of CompMLSM or MLSM can compute this meo, and the corresponding minimal triangulation is not obtainable by any of these algorithms.*

Proof. Let $G = (V, E)$ be a graph, and let S be a labeling structure. Clearly, any S-MLSM minimal triangulation of G is an S-CompMLSM one.

Conversely, let H be an S-CompMLSM minimal triangulation of G, and let us show that it is an S-MLSM one. Let α be the ordering on V computed by some execution of S-CompMLSM computing H, and, for any i from 1 to n, let C_i be the connected component of G'_i chosen at Step i of this execution. Let α' be the ordering on V and H' be the minimal triangulation of G computed by an execution of S-MLSM, choosing, for any i from 1 to n, $\alpha'(i)$ at Step i in the following way (every variable v is denoted v in the execution of CompMLSM and v' in that of MLSM):

(1) Choose a connected component C'_i of G''_i containing a vertex of maximal label in G''_i.

(2) If there is some j from 1 to n such that $C'_i = C_j$ and $label'_i(\alpha(j))$ is maximum in C'_i, then choose $\alpha'(i) = \alpha(j)$, otherwise choose $\alpha'(i)$ equal to any vertex of C'_i of maximal label in G''_i.

Note that there is at most one integer j such that $C'_i = C_j$ since for any j, k such that $j < k$, $C_j \neq C_k$, since $\alpha(k) \in C_k \setminus C_j$. Let us show that $H' = H$. For any subset J of $\{1, 2, \ldots, n\}$, let $\alpha(J)$ denote the set of vertices $\{\alpha(j) \mid j \in J\}$, and, for any i from 1 to n, let $P(i)$ be the following property:

$P(i)$: If there is some j from 1 to n such that $C'_i = C_j$ and $\forall y \in C'_i$, $\alpha'(Num'_i(y)) = \alpha(Num_j(y))$, then the edges of H' produced when processing the vertices of C'_i (in the execution of MLSM) are exactly those of H produced when processing the vertices of C_j (in the execution of CompMLSM).

Let us show $P(i)$ by induction on i from 1 to n. $P(1)$ holds since C'_1 contains a single vertex and processing this vertex produces no edge of H (or H'). We suppose $P(i-1)$ for some i, $1 < i \leq n$. Let us show $P(i)$. We suppose that there is some j from 1 to n such that $C'_i = C_j$ and $\forall y \in C'_i$, $\alpha'(Num'_i(y)) = \alpha(Num_j(y))$. By Lemma 4.14, $\forall y \in C_j$, $Num_j(y) \subseteq Num_j(\alpha(j))$, so $\forall y \in C'_i$, $\alpha'(Num'_i(y)) = \alpha(Num_j(y)) \subseteq \alpha(Num_j(\alpha(j))) = \alpha'(Num'_i(\alpha(j)))$, and therefore $Num'_i(y) \subseteq Num'_i(\alpha(j))$. It follows by Lemma 4.4 that $label'_i(\alpha(j))$ is maximum in C'_i. By definition of α', $\alpha'(i) = \alpha(j)$. The edges of H' produced when processing $\alpha'(i)$ are exactly those of H produced when processing $\alpha(j)$ by Lemma 4.5(a) and the fact that $\forall y \in C'_i$, $\alpha'(Num'_i(y)) = \alpha(Num_j(y))$, and the connected components of G''_{i-1} obtained from C'_i by removing $\alpha'(i)$ are exactly those of G'_{j-1} obtained from C_j by removing $\alpha(j)$, with $\forall y \in C'_i \setminus \{\alpha'(i)\}$, $\alpha'(Num'_{i-1}(y)) = \alpha(Num_{j-1}(y))$. For each such connected component C, there is some $k < i$ and some $l < j$ such that $C = C'_k = C_l$ and $\forall y \in C'_k$, $\alpha'(Num'_k(y)) = \alpha'(Num'_{i-1}(y)) = \alpha(Num_{j-1}(y)) = \alpha(Num_l(y))$, so

by induction hypothesis, the edges of H' produced when processing the vertices of C are exactly those of H produced when processing the vertices of C. Hence, the edges of H' produced when processing the vertices of C'_i are exactly those of H produced when processing the vertices of C_j. So $P(i)$ holds, which completes the induction on i. Now, for each connected component C of G, there are some i and j from 1 to n such that $C = C'_i = C_j$ and $\forall y \in C$, $Num'_i(y) = Num_j(y) = \emptyset$, so by $P(i)$, the edges of H' produced when processing the vertices of C are exactly those of H produced when processing the vertices of C. Hence, $H' = H$. □

Theorem 4.16, together with Theorem 4.15, yields the following interesting result.

THEOREM 4.17. *For any graph G, whichever meo-computing algorithm of the MLSM and CompMLSM families is used (e.g., LEX M, MCS-M, LexDFS-M, MNSM, or their Comp extensions), the set of computable minimal triangulations is the same.*

These minimal triangulations fail to cover all possible minimal triangulations: Figure 7(b) shows a minimal triangulation which is obtainable by none of our graph search meo-computing algorithms.

5. Complexity of MLS and MLSM. We now consider the question of implementing Algorithms MLS and MLSM. In the following, we use the word "implementation" in an algorithmic sense and not in a programming one. We will give a detailed version of each algorithm studied in this paper and precise the data structures used in order to compute its time and space complexity, but we will not give any real implementation in a programming language. We will also explore variants Test-MLS and Test-MLSM: Algorithm Test-MLS takes as input a graph $G = (V, E)$, a labeling structure S, and an ordering α of V and returns "YES" if α is an S-MLS ordering of G and "NO" otherwise. It is obtained from Algorithm MLS by replacing the instructions

Choose a vertex x of G' of maximal label;
$\alpha(i) \leftarrow x$

by

if the label of $\alpha(i)$ is not maximal in G', then return "NO"
$x \leftarrow \alpha(i)$

and by adding the instruction

return "YES"

at the end of the algorithm.

Test-MLSM is defined from MLSM in the same way, and we denote by Test–S-MLS, Test–S-MLSM, Test-LexBFS, etc., the corresponding variants of S-MLS, S-MLSM, LexBFS, etc.

An implementation of S-MLS is required to compute only S-MLS orderings but not to be able to compute every S-MLS ordering of a graph. For instance, as any MCS ordering is an MNS ordering, any implementation of MCS is also an implementation of MNS. However, an implementation finding out if a given ordering is an MCS ordering or not will not be able to find out if this ordering is an MNS ordering or not. An implementation of Test–S-MLS (resp. Test–S-MLSM) will, in general, have to keep closer to S than S-MLS (resp. S-MLSM). By Lemmas 3.5 and 4.11, we have the following property.

PROPERTY 5.1. *Let S and S' be labeling structures with partial orders $\preceq_S$ and $\preceq_{S'}$, respectively, such that for any subsets I and I' of $\mathbb{Z}_2^+$, if $lab_{S'}(I) \prec_{S'} lab_{S'}(I')$, then $lab_S(I) \prec_S lab_S(I')$.*

Then any implementation of S-MLS (resp. S-MLSM) is also an implementation of S'-MLS (resp. S'-MLSM).

In particular, S-MLS (resp. S-MLSM) can be implemented by replacing the partial order on labels by any one of its linear extensions.

5.1. Complexity of MLS and Test-MLS. We will first study the main instances of MLS, namely, LexBFS, MCS, LexDFS, and MNS. Then we will give an implementation of S-MLS with a stratified tree, which is a data structure designed to manipulate priority queues, for any labeling structure S such that labels are positive integers ordered by $\leq$.

An implementation of LexBFS is given by Rose, Tarjan, and Lueker [14], and an implementation of MCS is given by Tarjan and Yannakakis [16] with the following complexity results.

THEOREM 5.2 (see [14, 16]). *LexBFS and MCS can be implemented in $O(n+m)$ time and space.*

It is easy to check that the data structures used in [14] for LexBFS and in [16] for MCS can be used without extra cost for Test-LexBFS and Test-MCS, respectively. Moreover, as by Theorem 3.6 any LexBFS or MCS ordering is an MNS one, we have the following Corollary of Theorem 5.2.

COROLLARY 5.3. *Test-LexBFS, Test-MCS, and MNS can be implemented in $O(n+m)$ time and space.*

We can derive from the implementation of LexBFS given in [14] an implementation of LexDFS.

Implementation of LexDFS and Test-LexDFS with a list of lists.

As in the implementation of LexBFS, the current state of labels is represented by a list L of nonempty lists l. Each list l contains the unnumbered vertices bearing a given label, and the list L is ordered in decreasing order on the labels associated with the lists l (see [14] for a full description of the data structure). It is initialized with a unique list l containing all the vertices of the graph. At Step i, $\alpha(i)$ is chosen in the first list in L and removed from this list, and for each list l in L, the neighbors of $\alpha(i)$ in l are removed from l and put into a new list l_1, which is placed just before l in L. This corresponds to the new decreasing LexBFS order in an execution of LexBFS. To obtain an implementation of LexDFS, it is sufficient to add the following instruction at the end of each Step i: scan the list L to extract the lists l_1 and form a list L_1 with these lists l_1 in the same order, then concatenate L_1 with the remaining list L to obtain the new list L in decreasing LexDFS order.

For Test-LexDFS, we test at iteration i whether $\alpha(i)$ is in the first list of L or not.

THEOREM 5.4. *LexDFS and Test-LexDFS can be implemented in $O(n^2)$ time and $O(n+m)$ space.*

Proof. The time complexity of LexDFS is obtained from the time complexity of LexBFS by adding the cost of scanning the list L to form the list L_1 at each step. As there are n scans and each scan requires $O(n)$ time, we obtain an $O(n^2)$ time complexity for LexDFS. The space complexity is the same as that of LexBFS, i.e., $O(n+m)$. These complexity bounds also hold for Test-LexDFS. □

We will now discuss implementing Test-MNS. We remark that for any vertices v and w of G'_i, the Boolean value of $label_{i-1}(v) \preceq label_{i-1}(w)$ depends only on the value of $label_i(v) \preceq label_i(w)$ and on whether v and w are neighbors of $\alpha(i)$ or not. Thus we can implement Test-MNS by storing these Boolean values instead of storing and comparing explicit labels.

Implementation of Test-MNS.

We use a Boolean matrix *Preceq* such that, at the beginning of Step i, for any vertices v and w of G'_i, $Preceq(v,w) = True$ iff $label_i(v) \preceq label_i(w)$. *Preceq* is

initialized with True. Testing the maximality of the label of $\alpha(i)$ in G' is implemented by testing the absence of a vertex v of G' such that $Preceq(\alpha(i), v)$ and not $Preceq(v, \alpha(i))$. Labels are updated at Step i by the following procedure, where $x = \alpha(i)$:

foreach neighbor v of x in G' **do**
  foreach non-neighbor w of x in G' **do**
    $Preceq(v, w) \leftarrow False$.

It is easy to check that this procedure correctly updates matrix $Preceq$ with respect to its desired meaning, so that this implementation of Test-MNS is correct.

THEOREM 5.5. *Test-MNS can be implemented in $O(n(n+m))$ time and $O(n^2)$ space.*

Proof. Initialization requires $O(n^2)$ time and at each Step i, testing the maximality of the label of $\alpha(i)$ in G' requires $O(n)$ time, and updating matrix $Preceq$ requires $O(n|N(\alpha(i))|)$ time, which makes a global $O(n(n+m))$ time bound. Matrix $Preceq$ requires $O(n^2)$ space, which is the space bound of the algorithm. ∎

Note that we can derive from this implementation of Test-MNS an implementation of MNS which is able to compute every MNS ordering of the input graph with the same time and space bounds. In addition to matrix $Preceq$, we use an array containing for each vertex v of G' the number of vertices w of G' having a larger label than v (i.e., such that $Preceq(v, w)$ and not $Preceq(w, v)$). This array allows us to choose at each step a vertex of maximal label in G' in $O(n)$ time.

5.1.1. Using a stratified tree. For any labeling structure S such that labels are positive integers ordered by $\leq$, S-MLS can be implemented with the data structure of a stratified tree defined by van Emde Boas [17, 18] to manipulate priority queues. This data structure is used to implement a subset C of an interval of integers in the form $[1, n]$ ordered by $\leq$, which here will be the set of current labels, i.e., the set of labels assigned to unnumbered vertices at some point of the execution of S-MLS. The stratified tree can be initialized in $O(n \log\log n)$ time. Inserting or removing an element, or finding the maximum element in C requires $O(\log\log n)$ time. The structure requires $O(n)$ space. These bounds are computed in the model of the unit-cost RAM.

Implementation of S-MLS and Test–S-MLS with a stratified tree.

We suppose that labels are positive integers in some interval I ordered by $\leq$. The set C of current labels is stored in a stratified tree. With each current label is associated the nonempty list of unnumbered vertices having this label. These lists can be stored in an array indexed by the integers in I. At the initialization step, the unique element l_0 of C is associated with the list of all vertices of the graph. To choose an unnumbered vertex with maximal label at Step i, we find the maximum element $l_{\max}$ of C, remove a vertex from the list associated with $l_{\max}$, and remove $l_{\max}$ from C if this list has become empty. For Test–S-MLS, it is tested whether $\sigma(i)$ has label $l_{\max}$ or not. To update the label of a vertex v from l to l', we transfer v from the list associated with l to the list associated with l', with a possible removal of l from C and a possible insertion of l' into C.

PROPOSITION 5.6. *Let S be a labeling structure such that for any integer $n \geq 1$, set $L_n = \{lab_S(I),\ I \subseteq [1, n]\}$ is a subset of an interval $[1, r(n)]$ of integers ordered by $\leq$, with $r(n)$ in $O(n)$.*

Then S-MLS and Test–S-MLS can be implemented in $O((n+m)\log\log n + m\, t_{Inc}(n))$ time and $O(n+m)$ space, where $t_{Inc}(n))$ is the time required to increment a label of L_n.

Proof. Implementing the set C of current labels with a stratified tree requires $O(n \log\log n)$ initialization time, $O(\log\log n)$ time per operation, and $O(n)$ space [17, 18]. As choosing a vertex with maximal label and updating the label of a vertex require at most a constant number of operations on the stratified tree, we obtain the announced bounds. □

As for CompMLS, for any labeling structure S, S-CompMLS can be implemented by any implementation of S-MLS, since any S–MLS ordering is an S-CompMLS one. Moreover, by Theorem 3.7, S-CompMLS restricted to chordal graphs can be implemented by an implementation of LexBFS or MCS. Test–S-CompMLS restricted to chordal graphs only has to determine if the given ordering is a peo of the graph or not, which can be implemented in linear time and space [14].

5.2. Complexity of MLSM and Test-MLSM. MLSM is more complex than MLS, since graph G' must additionally be searched at each Step i to determine the vertices whose label has to be incremented, i.e., the neighbors of $\alpha(i)$ in the minimal triangulation H of G. We will show that this search can be performed using another labeling structure than the input labeling structure S, which will allow us to deduce an implementation and complexity bounds of S-MLSM from those of LEX M and S-MLS.

An implementation of LEX M is given by Rose, Tarjan, and Lueker [14] with the following complexity results, where m' denotes the number of edges of the computed minimal triangulation H of G.

THEOREM 5.7 (see [14]). *LEX M can be implemented in $O(n(n+m))$ time and $O(n+m')$ space.*

Note that the implementation of LEX M given in [14] requires only $O(n+m)$ space because it only computes an meo α of G and not its minimal triangulation $H = G^+_\alpha$. We will show that this implementation of LEX M can be used to implement S-MLS for any labeling structure S. We first extend Lemma 4.5(a) to the following lemma.

LEMMA 5.8. *For any graph G, any labeling structures S and S' with partial orders $\preceq_S$ and $\preceq_{S'}$, respectively, any execution of MLSM on G and S, any integer i between 1 and n, and any path μ in G'_i ending in some vertex y, $\forall t \in \mu \setminus \{y\}$, $label_i(t) \prec_S label_i(y)$ iff $\forall t \in \mu \setminus \{y\}$, $lab_{S'}(Num_i(t)) \prec_{S'} lab_{S'}(Num_i(y))$.*

Proof. By Lemma 4.5(a), it is sufficient to show that $\forall t \in \mu \setminus \{y\}$, $lab_{S'}(Num_i(t)) \prec_{S'} lab_{S'}(Num_i(y))$ iff $\forall t \in \mu \setminus \{y\}$, $Num_i(t) \subset Num_i(y)$. The proof of this last equivalence is similar to that of Lemma 4.5(a), replacing the references to Lemma 4.4 with references to condition IC. □

We distinguish in an execution of MLSM the *specific MLSM part*, which is the search in G' at each Step i from $\alpha(i)$ to determine the vertices whose label has to be incremented, from the *MLS part* which corresponds to an execution of MLS on the output graph G^+_α. We suppose in the following result that the MLS part of MLSM on G can be implemented with the same time and space complexity as MLS on G^+_α. In other words, we assume that the fact that G^+_α is only partially known at each step of an execution of MLSM does not affect the complexity of MLS, which seems reasonable since the unknown edges of G^+_α are between unnumbered edges and therefore play no role in the algorithm.

THEOREM 5.9. *For any labeling structure S, if S-MLS (resp. Test–S-MLS) can be implemented in $O(f(n,m))$ time and $O(g(n,m))$ space, then S-MLSM (resp. Test–S-MLSM) can be implemented in $O(n(n+m)+f(n,m'))$ time and $O(g(n,m'))$ space.*

Proof. We implement the specific MLSM part of Algorithm S-MLSM or Test-S-MLSM with the part of the implementation of LEX M given in [14] corresponding to the specific MLSM part and the updating of the integer labels $l(v)$ at each step. As the order on these labels $l(v)$ at the beginning of Step i is the same as the lexicographic

order on the $lab_{S_1}(Num_i(v))$, with S_1-MLSM = LEX M [14], Lemma 5.8 ensures that this correctly implements the specific MLSM part of the algorithm. By our assumption, the MLS part of S-MLSM (resp. Test–S-MLSM) can be implemented by an implementation of S-MLS (resp. Test–S-MLS) with the same complexity as this algorithm on G^+_α. The result follows from Theorem 5.7. □

COROLLARY 5.10. *MCS-M, LexDFS-M, MNSM, Test-LEX M, Test-MCS-M, and Test-LexDFS-M can be implemented in $O(n(n+m))$ time and $O(n+m')$ space.*

Test-MNSM can be implemented in $O(n(n+m'))$ time and $O(n^2)$ space.

For some labeling structures S, we can directly derive from the implementation of LEX M given in [14] an implementation of S-MLSM by just modifying the way labels $l(v)$ are updated to make them correspond to the labels obtained with the labeling structure S. For instance, we obtain an implementation of LexDFS-M by replacing the instruction $l(z) := l(z) + 1/2$ by $l(z) := l(z) + n$, and we obtain an implementation of MCS-M by replacing this instruction by $l(z) := l(z) + 1$ and by replacing the procedure *sort* by the following: set k to the maximum value of $l(v)$ for unnumbered vertices v. This implementation of MCS-M is a little simpler than the implementation of LEX M (with the same complexity bounds), since it avoids renaming all labels in the procedure *sort*. It can be used in the implementation of S-MLSM and Test–S-MLSM instead of the implementation of LEX M and is itself an implementation of MNSM by Theorem 4.12. In the same way, we can define direct implementations of Test-LexDFS-M and Test-MCS-M, but not of Test-MNSM. If labels are positive integers ordered by $\leq$, we can implement the MLS part with a stratified tree and deduce complexity bounds from Proposition 5.6 and Theorem 5.9. By Theorem 4.15, for any labeling structure S, S-CompMLSM can be implemented by an implementation of LEX M or MCS-M.

6. Conclusion. We have extended Algorithm LexBFS into Algorithm MLS by defining a general labeling structure and shown how to extend this further to CompMLS to enable any possible peo to be computed. We have also extended all these algorithms to meo-computing versions. Our work yields alternate (and often simpler) proofs for the results of several papers, as [1, 14, 15, 16, 19].

However, we have shown that these new meo-computing algorithms fail to enhance the possibility for finding a wider range of minimal triangulations. LEX M has been studied experimentally and shown to be very restrictive (see [5]), yielding triangulations which, for example, are far from edge-number minimum. This problem remains with the enlarged family of new meo-computing algorithms we present here and appears to be a fundamental limitation of graph search.

We presented time and space complexity bounds of some algorithms of the MLS and MLSM families. These results mostly derive from the known complexity bounds of LexBFS, MCS, and LEX M. An interesting fact is that the search in G' at each step of an execution of MLSM can be performed using any other labeling structure than the input labeling structure S, which allows us to implement S-MLSM by combining implementations of LEX M and S-MLS.

As mentioned in the Introduction, LexBFS and MCS, though designed for chordal graphs, have been used for graph classes other than chordal graphs. The more general peo-finding algorithms discussed in this paper could also prove useful on non-chordal graphs, on a wider variety of graph classes and problems.

Appendix. We give the proof of Lemma 4.5.

LEMMA 4.5. *For any graph G, any execution of MLSM on G computing ordering α, any integer i from 1 to n, and any path μ in G'_i ending in some vertex y,*

(a) $\forall t \in \mu \setminus \{y\},\ label_i(t) \prec label_i(y)$ *iff* $\forall t \in \mu \setminus \{y\},\ Num_i(t) \subset Num_i(y)$;
(b) *if* $\forall t \in \mu \setminus \{y\},\ label_i(t) \prec label_i(y)$, *then* $\forall t \in \mu \setminus \{y\},\ \alpha^{-1}(t) < \alpha^{-1}(y)$;
(c) *if* $\forall t \in \mu \setminus \{y\},\ \alpha^{-1}(t) < \alpha^{-1}(y)$, *then* $\forall t \in \mu \setminus \{y\},\ Num_i(t) \subseteq Num_i(y)$.

The proof uses the following technical Lemmas 6.1 and 6.2. For any path μ containing vertices x and y, $\mu[x, y]$ denotes the subpath of μ between x and y.

LEMMA 6.1. *For any graph G, any execution of MLSM on G, any integer i from 1 to n, and any path μ in G'_{i-1} ending in some vertex y, if $\forall t \in \mu \setminus \{y\},\ Num_i(t) \subset Num_i(y)$, then $\forall t \in \mu \setminus \{y\},\ Num_{i-1}(t) \subset Num_{i-1}(y)$.*

Proof. We suppose that $\forall t \in \mu \setminus \{y\},\ Num_i(t) \subset Num_i(y)$ (and, therefore $label_i(t) \prec label_i(y)$ by Lemma 4.4). Let $t \in \mu \setminus \{y\}$, and let us show that $Num_{i-1}(t) \subset Num_{i-1}(y)$. It is sufficient to show that if $i \in Num_{i-1}(t)$, then $i \in Num_{i-1}(y)$. We suppose that $i \in Num_{i-1}(t)$. By Lemma 4.3, there is a path λ in G'_i from $\alpha(i)$ to t such that $\forall t' \in \lambda \setminus \{\alpha(i), t\},\ label_i(t') \prec label_i(t)$. Let μ' be the path obtained by the concatenation of λ and $\mu[t, y]$. Then μ' is a path in G'_i from $\alpha(i)$ to y such that $\forall t' \in \mu' \setminus \{\alpha(i), y\},\ label_i(t') \prec label_i(y)$. Hence, by Lemma 4.3, $i \in Num_{i-1}(y)$. ∎

LEMMA 6.2. *For any graph G, any execution of MLSM on G, any integer i from 1 to n, and any path μ in G'_i ending in some vertex y, if $\exists t \in \mu \setminus \{y\} \mid Num_i(t) \not\subseteq Num_i(y)$, then $\exists t_1 \in \mu \setminus \{y\} \mid \forall t \in \mu[t_1, y] \setminus \{t_1\},\ Num_i(t) \subset Num_i(t_1)$.*

Proof. We suppose that $\exists t \in \mu \setminus \{y\} \mid Num_i(t) \not\subseteq Num_i(y)$. Let j be the largest integer such that $\exists t \in \mu \setminus \{y\} \mid Num_{j-1}(t) \not\subseteq Num_{j-1}(y)$, and let t_1 be the vertex of μ closest to y such that $Num_{j-1}(t_1) \not\subseteq Num_{j-1}(y)$. So $j - 1 \geq i$, $j \in Num_{j-1}(t_1)$, and $\forall t \in \mu[t_1, y] \setminus \{t_1\},\ j \notin Num_{j-1}(t)$. Let us show that $Num_j(t_1) = Num_j(y)$. We assume for contradiction that $Num_j(t_1) \neq Num_j(y)$. Let t_2 be the vertex of $\mu[t_1, y]$ closest to t_1 such that $Num_j(t_2) = Num_j(y)$. By the choice of j, $\forall t \in \mu[t_1, t_2] \setminus \{t_2\},\ Num_j(t) \subset Num_j(t_2)$, and, by Lemma 6.1, $Num_{j-1}(t_1) \subset Num_{j-1}(t_2)$. So $j \in Num_{j-1}(t_2)$, with $t_2 \in \mu[t_1, y] \setminus \{t_1\}$, a contradiction.

So $\forall t \in \mu[t_1, y] \setminus \{t_1\},\ Num_{j-1}(t) = Num_j(t) \subseteq Num_j(y) = Num_j(t_1) \subset Num_j(t_1) \cup \{j\} = Num_{j-1}(t_1)$. As $j - 1 \geq i$, by Lemma 6.1, $\forall t \in \mu[t_1, y] \setminus \{t_1\},\ Num_i(t) \subset Num_i(t_1)$. ∎

Proof of Lemma 4.5. (a) For the forward direction, we assume for contradiction that $\forall t \in \mu \setminus \{y\},\ label_i(t) \prec label_i(y)$, and $\exists t \in \mu \setminus \{y\} \mid Num_i(t) \not\subset Num_i(y)$ (and, therefore $Num_i(t) \not\subseteq Num_i(y)$, since, by Lemma 4.4 (i), $Num_i(t) \neq Num_i(y)$). By Lemma 6.2, $\exists t_1 \in \mu \setminus \{y\} \mid Num_i(y) \subset Num_i(t_1)$, and, by Lemma 4.4, $label_i(y) \prec label_i(t_1)$, a contradiction.

The reverse direction follows immediately from Lemma 4.4.

(b) We suppose that $\forall t \in \mu \setminus \{y\},\ label_i(t) \prec label_i(y)$. Let $k = \max\{\alpha^{-1}(t),\ t \in \mu\}$. By (a) and Lemma 6.1, $\forall t \in \mu \setminus \{y\},\ label_k(t) \prec label_k(y)$. So $\alpha(k) = y$, which completes the proof.

(c) We assume for contradiction that $\forall t \in \mu \setminus \{y\},\ \alpha^{-1}(t) < \alpha^{-1}(y)$ and $\exists t \in \mu \setminus \{y\} \mid Num_i(t) \not\subseteq Num_i(y)$. By Lemma 6.2, $\exists t_1 \in \mu \setminus \{y\} \mid \forall t \in \mu[t_1, y] \setminus \{t_1\},\ Num_i(t) \subset Num_i(t_1)$, and by (a) and (b), $\alpha^{-1}(y) < \alpha^{-1}(t_1)$, a contradiction. ∎

REFERENCES

[1] A. BERRY, J. BLAIR, P. HEGGERNES, AND B. PEYTON, *Maximum cardinality search for computing minimal triangulations of graphs,* Algorithmica, 39 (2004), pp. 287–298.
[2] A. BERRY AND J.-P. BORDAT, *Separability generalizes Dirac's theorem,* Discrete Appl. Math., 84 (1998), pp. 43–53.

[3] A. Berry and J.-P. Bordat, *Moplex elimination orderings,* in Proceedings of the First Cologne-Twente Workshop on Graphs and Combinatorial Optimization, Electron. Notes Discrete Math. 8, J. Hurink, S. Pickl, H. Broersma, and U. Faigle, eds., Elseiver, Amsterdam, 2001, pp. 6–9.

[4] A. Berry, R. Krueger, and G. Simonet, *Ultimate generalizations of LexBFS and LEX M,* in Proceedings of the 31st International Workshop on Graph-Theoretic Concepts in Computer Science 2005 (WG 2005), Lect. Notes Comput. Sci. 3787, Springer-Verlag, New York, 2005, pp. 199–213.

[5] J. R. S. Blair, P. Heggernes, and J. A. Telle, *A practical algorithm for making filled graphs minimal,* Theoret. Comput. Sci. A, 250-1/2 (2001), pp. 125–141.

[6] H. L. Bodlaender and A. M. C. A. Koster, *On the maximum cardinality search lower bound for treewidth,* Discrete Appl. Math., 155 (2007), pp. 1348–1372.

[7] D. G. Corneil, *Lexicographic breadth first search—a survey,* in Proceedings of the 30th International Workshop on Graph Theory (WG2004), Lect. Notes Comput. Sci. 3353, Springer-Verlag, New York, 2004, pp. 1–19.

[8] D. G. Corneil and R. M. Krueger, *Unified view of graph searching,* SIAM J. Discrete Math., 22 (2008), pp. 1259–1276.

[9] D. G. Corneil, S. Olariu, and L. Stewart, *Linear time algorithms for dominating pairs in asteroidal triple-free graphs,* SIAM J. Comput., 28 (1999), pp. 1284–1297.

[10] E. Dahlhaus, P. L. Hammer, F. Maffray, and S. Olariu, *On Domination Elimination Orderings and Domination Graphs,* in Proceedings of WG 1994, London, 1994, pp. 81–92.

[11] D. R. Fulkerson and O. A. Gross, *Incidence matrixes and interval graphs,* Pacific J. Math., 15 (1965), pp. 835–855.

[12] T. Ohtsuki, *A fast algorithm for finding an optimal ordering in the vertex elimination on a graph,* SIAM J. Comput., 5 (1976), pp. 133–145.

[13] T. Ohtsuki, L. K. Cheung, and T. Fujisawa, *Minimal triangulation of a graph and optimal pivoting order in a sparse matrix,* J. Math. Anal. Appl., 54 (1976), pp. 622–633.

[14] D. Rose, R. E. Tarjan, and G. Lueker, *Algorithmic aspects of vertex elimination on graphs,* SIAM J. Comput., 5 (1976), pp. 266–283.

[15] D. R. Shier, *Some aspects of perfect elimination orderings in chordal graphs,* Discrete Appl. Math., 7 (1984), pp. 325–331.

[16] 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.

[17] P. van Emde Boas, *Preserving order in a forest in less than logarithmic time and linear space,* Inform. Process. Lett., 6 (1977), pp. 80–82.

[18] P. van Emde Boas, R. Kaas, and E. Zijlstra, *Design and implementation of an efficient priority queue,* Math. Syst. Theory, 10 (1977), pp. 99–127.

[19] Y. Villanger, *Lex M versus MCS-M,* Discrete Math., 306 (2004), pp. 393–400.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 447–465

MORE PATTERNS IN TREES: UP AND DOWN, YOUNG AND OLD, ODD AND EVEN*

NACHUM DERSHOWITZ† AND SHMUEL ZAKS‡

Abstract. We apply the tree-pattern enumeration formulæ of earlier work of ours [N. Dershowitz and S. Zaks, *Discrete Appl. Math.*, 25 (1989), pp. 241–255], and a new extension thereof, to some recent enumerations of distributions of leaves in ordered trees [W. Y. C. Chen, E. Deutsch, and S. Elizalde, *European J. Combin.*, 27 (2006), pp. 414–427] and in bicolored ordered trees [L. H. Clark, J. E. McCanna, and L. A. Székely, *Bull. Inst. Combin. Appl.*, 21 (1997), pp. 33–45], and of distributions of up-down-up subpaths in Dyck lattice paths [Y. Sun, *Discrete Math.*, 287 (2004), pp. 177–186]. Bijections are used to facilitate the derivation of statistics for bicolored trees.

Key words. tree enumerations, tree patterns, node distribution, ordered trees, plane-planted trees, bicolored trees, binary trees, Dyck paths, lattice paths, bridges

AMS subject classification. 05A15

DOI. 10.1137/070687475

the bridge guard's bucket
upside-down to dry...
fresh leaves
—Issa (1818)

1. Introduction. Over and above their intrinsic combinatorial interest, enumerations of classes of trees have manifold applications to average-case analysis of algorithms. For instance, the performance of various manipulations of tree structures may depend on the distribution of node degrees or of tree heights. For one example, see [12].

Several recent lattice-path and tree enumerations turn out to be amenable to the generic pattern enumeration formula we gave in [7], which was based on Dvoretsky and Motzkin's cycle lemma [10] (called "penetrating analysis" in [16]; see [8]). One formulation of this lemma is the following.

CYCLE LEMMA (see [10]). *For any sequence of m natural numbers $j_0, \ldots, j_{m-1}$, whose sum is n, with $m > n$, there are exactly $m - n$ offsets π, $0 \le \pi < m$, such that*

$$j_{\pi \bmod m} + \cdots + j_{(\pi+k-1) \bmod m} > k + n - m$$

for all k, $1 \le k < m$.

For example, the sequence 102001030 has 2 $(= 9 - 7)$ such cyclic permutations:

$$\underline{10}\ \underline{30\underline{10}\,\underline{200}} \qquad \underline{30\underline{10}\,\underline{200}}\ \underline{10}.$$

Notice that both of these sequences can be read as well-formed Polish-prefix expressions (as indicated by the underscoring), where each number j is followed by j well-formed expressions.

*Received by the editors April 4, 2007; accepted for publication (in revised form) July 28, 2008; published electronically January 16, 2009.
http://www.siam.org/journals/sidma/23-1/68747.html
†Department of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel (nachum.dershowitz@cs.tau.ac.il). This author's research was supported in part by the Israel Science Foundation (grant 250/05).
‡Department of Computer Science, Technion, Haifa 32000, Israel (zaks@cs.technion.ac.il).

To understand why the lemma holds, note that such a sequence must have at least one occurrence of 0, since $m > n$. Replacing a (cyclically adjacent) pair $j0$, where j is positive, with just $j-1$, decreases both the sum n and count m by 1. This does not affect the quantity—or starting positions—of valid cyclic permutations, which cannot begin with 0. Repeatedly making such replacements terminates when only zeros remain, at which point the lemma clearly holds true.

Specifically, we show here how Sun's enumeration [23] of Dyck lattice paths with a given number of subsequences $\nearrow\searrow\nearrow$ (**udu**) can be solved in this manner, and how a convenient extension of our pattern formula applies to the enumeration by Chen, Deutsch, and Elizalde [4] of ordered trees with given numbers of eldest and noneldest childless children (leaves). We also show how to apply the extended formula to count trees with given distributions of leaves on odd and even levels, as done by Clark, McCanna, and Székely [5].

We review the pattern enumeration of [7] in the next two sections, and then extend it in section 4. Our main result, Theorem 4.1 below, counts occurrences of a multiset of patterns in ordered trees, or in other Catalan structures (see Figure 1). Tree patterns are specified by subtrees having empty slots that can be filled with leaves and/or with larger subtrees, and gaps that can be filled by a series of subtrees (see Figure 2).

These formulæ are applied to the problems of Sun [23], of Chen, Deutsch, and Elizalde [4], and of Clark, McCanna, and Székely [5] in sections 5, 6, and 7, respectively. Section 7 includes a new bijection between ordered trees, mapping odd-level nodes to internal nodes, and even-level internal nodes to leftmost leaves, plus some results on their distributions.

In the concluding section, the correspondence between the lattice-path enumeration of [23], the ordered-tree enumeration of [4], and the bicolored tree enumeration of [5] is explicated.

2. Pattern formulæ. *Ordered trees* (that is, plane-planted trees with a root whose children are also ordered in sequence) may be defined as follows:

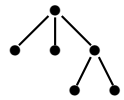
$$T ::= \langle T^* \rangle \,,$$

meaning a bracketed sequence of zero or more ordered trees.[1] In other words, T is the set of nonempty *balanced* bracketed expressions, where the first open bracket matches the *last* close bracket. See Figure 1(a). A *subtree* of $t \in T$ is any subsequence of t that is itself a tree in T. Ordered trees are counted by the Catalan/Segner [2, 20] numbers, and are in one-to-one correspondence with many other combinatorial structures, including those in Figure 1(b),(c). See [13].

Patterns come in four basic shapes: $\triangle$, $\Diamond$, $\blacktriangle$, and $\infty\!\!\infty$. The plain triangle *pattern* $\triangle$ matches any subtree, a lozenge $\Diamond$ corresponds to any tree leaf $\langle\rangle$, a dark triangle $\blacktriangle$ matches any nonleaf subtree (that is, a subtree rooted at an internal node), and an ellipsis $\infty\!\!\infty$ can match any sequence of (zero or more) subtrees. So, the pattern $\triangle$ matches any subtree matched by either $\Diamond$ or $\blacktriangle$.

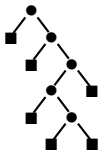
Base patterns can be composed to form more complicated shapes. An ellipsis is intended to match a forest (sequence) of trees, rather than a single tree, so it makes sense only within a composite pattern. Thus, tree patterns have the following

[1] Some authors (e.g., [24, 5]) follow a convention by which ordered trees have an extra unary (monovalent) node connected by an extra edge to what is the root of our trees. This necessitates minor changes in the parameters of some enumerations.



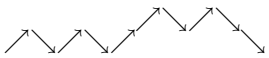
$\langle\langle\rangle\,\langle\rangle\,\langle\langle\rangle\,\langle\rangle\rangle\rangle$

(a) Ordered tree



$\langle\langle\rangle\,\langle\langle\rangle\,\langle\langle\langle\rangle\,\langle\langle\rangle\,\langle\rangle\rangle\rangle\,\langle\rangle\rangle\rangle\rangle$

(b) Binary tree



ududuududd

(c) Lattice path

FIG. 1. *Corresponding objects.*

grammar:

$$P ::= \Diamond \mid \triangle \mid \blacktriangle \mid \langle Q^+ \rangle\,,$$
$$Q ::= P \mid \infty\infty\,,$$

where Q^+ means one or more patterns Q, in sequence. See Figure 2.

A *composite* pattern $\langle p_1 \cdots p_q \rangle$ *matches* a tree $\langle t_1 \cdots t_n \rangle$ if the latter's immediate subtrees $t_1 \cdots t_n$ can be divided into q (possibly empty) subsequences $t_1 \cdots t_{k_1} \cdots t_{k_2} \cdots t_{k_q} = t_n$ of trees, such that each p_i $(i = 1, \ldots, q)$ matches the subsequence $t_{k_{i-1}+1} \cdots t_{k_i}$ $(k_0 = 0)$. Of course, only the ellipsis pattern $\infty\infty$ can match a subsequence of more than one subtree; an ellipsis even matches the empty sequence of zero subtrees. In other words, a match is an injection (embedding) of pattern nodes to tree nodes and of pattern edges to tree edges that preserves edge incidence and order, and also edge neighbors—unless the pattern has an ellipsis between the edges in question.

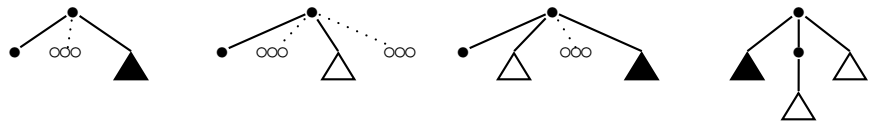
For example, the pattern $\langle \Diamond\, \infty\infty\, \blacktriangle \rangle$, depicted in Figure 2(a), matches any tree whose root has two or more children, the youngest (rightmost) having children, but the eldest (leftmost) still childless (a leaf). Thus, it matches the tree t, depicted in Figure 1(a). Clearly, the same tree can match many different (base and composite) patterns. In fact, t is also matched by the patterns in Figure 2(b),(c) but not by that in Figure 2(d).

There may be many ways to divide n subtrees into q subsequences when a pattern has more than one ellipsis. For instance, the pattern $\langle \Diamond \infty\infty\, \triangle \infty\infty \rangle$ in Figure 2(b) matches t in two ways, for each of the two younger children.

A pattern *occurs* in a tree if it matches any *sub*tree; for example, $p = \langle \Diamond \infty\infty\, \triangle \infty\infty \rangle$ (Figure 2(b)) also occurs at the youngest child of the root of t (Figure 1(a)). A multiset (bag) of patterns occurs in a tree if the patterns match *disjoint* subtrees. The singleton set $\{\langle \Diamond \Diamond \rangle\}$ of patterns appears exactly once in t, while $\{\Diamond, \Diamond\}$ appears six times, once for each choice of two of the leaves. The disjointness requirement means that no tree node or edge may be part of more than one pattern match—though triangles (plain $\triangle$ or dark $\blacktriangle$) of one pattern can be at the root of an occurrence of another pattern. Thus, the pattern multiset $\{p, p\}$ occurs only twice in t: one p at the youngest child and the other at the root in either of two ways.

The number of edges in a pattern is equal to the total number of left (or right) brackets—excluding the first one, plus the total number of triangles, $\triangle$ or $\blacktriangle$, plus the total number of leaf patterns $\Diamond$. In other words, the number of edges $e(p)$ in a (nonellipsis) pattern p is obtained, recursively, as follows:

$$e(\triangle) = e(\blacktriangle) = e(\Diamond) = 0,$$



$\langle \Diamond \circ\circ\circ \blacktriangle \rangle$ $\langle \Diamond \circ\circ\circ \triangle \circ\circ\circ \rangle$ $\langle \Diamond \triangle \circ\circ\circ \blacktriangle \rangle$ $\langle \blacktriangle \langle \triangle \rangle \triangle \rangle$

(a) Leftmost leaf; rightmost internal

(b) Leftmost leaf; another child

(c) Leftmost leaf; adjacent sibling; rightmost internal

(d) Ternary node; leftmost internal; middle unary

FIG. 2. *Some patterns.*

$$e(\circ\circ\circ) = -1,$$

$$e(\langle p_1 \cdots p_q \rangle) = e(p_1) + \cdots + e(p_q) + q.$$

The number $v(p)$ of nodes (vertices) in p can be calculated similarly:

$$v(\triangle) = v(\blacktriangle) = v(\circ\circ\circ) = 0,$$

$$v(\Diamond) = 1,$$

$$v(\langle p_1 \cdots p_q \rangle) = v(p_1) + \cdots + v(p_q) + 1.$$

For example, the four patterns in Figure 2 comprise a total of $e = 2 + 2 + 3 + 4 = 11$ edges and $v = 2 + 2 + 2 + 2 = 8$ nodes.

Let $p_1, \ldots, p_q$, be some patterns and let their desired multiplicities of occurrence in a tree be $n_1, \ldots, n_q$, respectively. In [7], we provided the pattern enumeration formula

$$\frac{1}{u} \binom{u}{n_1, \ldots, n_q, u-m} \binom{2n + d + s - 2e - m}{n - e} \tag{2.1}$$

for the number of occurrences of such a multiset of $m = \Sigma n_i$ patterns among all n-edge ordered trees, where e is the total number of edges in the m patterns, d is the number of plain triangles $\triangle$ therein, s is the number of ellipses $\circ\circ\circ$, and $u = n + d - e + 1$. Nonleaf patterns $\blacktriangle$ were not treated in [7].[2]

This formula can be used to calculate various tree statistics. Let the *size* of a tree be measured by the number of its edges.

Example 2.1. A simple one-pattern case of this formula establishes the expected number of nodes in an ordered tree that are "eldest" leaves (leftmost and childless). The pattern $p_1 = \langle \Diamond \circ\circ\circ \rangle$ matches each such leaf. Letting $m = n_1 = e = s = 1$, $d = 0$, $u = n$ in (2.1), yields

$$\frac{1}{n} \binom{n}{1} \binom{2n-2}{n-1} = \binom{2n-2}{n-1}$$

[2]We use multinomial coefficients $\binom{j}{j_1, \ldots, j_k} = \frac{j!}{j_1! \cdots j_k!}$ throughout. Virtually all the formulæ in this paper containing occurrences of the multinomial would benefit from dropping the common (but not universal) requirement that $j = j_1 + \cdots + j_k$, in which case the first two factors of (2.1) would become simply $\binom{u-1}{n_1, \ldots, n_q, u-m}$.

for the total number of leftmost leaves among all trees of size n. The number of trees of size n is the Catalan number [14],

$$C_n = \frac{1}{2n+1}\binom{2n+1}{n} = \frac{1}{n+1}\binom{2n}{n}$$

(see [21, A000108]); so there are, on the average,

$$\frac{\binom{2n-2}{n-1}}{\frac{1}{n+1}\binom{2n}{n}} = \frac{n^2+n}{4n-2} \approx \frac{n}{4}$$

leftmost leaves in a tree with $n+1$ nodes.

Similarly, the average number of "younger" (nonleftmost) leaves is counted by taking $p_1 = \langle \triangle \circ\!\!\circ \Diamond \circ\!\!\circ \rangle$, $m = n_1 = d = 1$, $e = s = 2$, and $u = n$:

$$\frac{\binom{2n-2}{n-2}}{\frac{1}{n+1}\binom{2n}{n}} = \frac{n^2-1}{4n-2} \approx \frac{n}{4}.$$

The leaf subpattern $\Diamond$ contributes to the count for *each* nonleftmost leaf child of the node matching $\langle \triangle \circ\!\!\circ \Diamond \circ\!\!\circ \rangle$. □

Here and throughout, we speak freely of nodes as being "leftmost," "nonleftmost," "rightmost," or "nonrightmost" interchangeably with age-based terminology, "oldest," "younger," "youngest," or "older," respectively. The root of a tree and the only child of a unary node are leftmost and rightmost at one and the same time.

The following is an easy observation.

PROPOSITION 2.2. *In any ordered tree, the number of leftmost (resp., rightmost) leaves is one more than the number of nonleftmost (resp., nonrightmost) internal nodes.*

The difference of one in the numbers is on account of the root, which is perforce leftmost, as well as rightmost.

This correspondence holds even for the one-node tree, which has one leftmost/rightmost leaf, the root, and no internal nodes at all. It also holds for the two-node tree, which has one leftmost/rightmost leaf and no nonleftmost internal nodes.

Proof. We give two proofs, one "geometric" and one "algebraic."

1. From each leftmost leaf, travel up leftmost edges as far as possible, reaching the root or else reaching some nonleftmost internal node. In the reverse, from the root or any nonleftmost internal node, one can travel down leftmost edges until encountering the corresponding leftmost leaf.

2. Suppose there are i internal nodes, ℓ leftmost leaves, and k leftmost nonroot internal nodes. Since every internal node has exactly one leftmost child, $i = \ell + k$. Hence $\ell = i - k$, which is one more (for the root) than the number of nonleftmost internal nodes.

(The correspondence of rightmost leaves with nonrightmost internal nodes and the root follows by symmetry.) □

We will see (Theorem 7.3 below) that leftmost leaves, rightmost leaves, leftmost/rightmost internal nodes, nonleftmost/nonrightmost leaves/internal nodes all occur with almost equal frequency among ordered trees with a given number of nodes (or of edges). All eight cases occur in about one-fourth of the nodes. See the previous example.

3. Tree enumerations. This paper is mainly concerned with counting trees, rather than pattern occurrences.

When the patterns are such that there can be no more than one occurrence of the multiset of patterns in any one tree, then formula (2.1) above counts trees. This is the case, in particular, when the patterns include all the nodes in the tree being counted ($m + e - d = n + 1$), and different patterns do not overlap each other. (Two patterns "overlap" if there is a tree in which both occur and whose occurrences share at least one node.) The latter condition precludes, for instance, sibling ellipses, such as $\langle \circ\circ\, \triangle \circ\circ \rangle$, which can occur multiple times at the same subtree.

Example 3.1. The number of ordered trees with n edges, ℓ leaves, and no unary or binary nodes is counted by the number of occurrences of $n_1 = \ell$ leaf patterns $\Diamond$ and $n_2 = n + 1 - \ell$ patterns $\langle \triangle\triangle\triangle\circ\circ \rangle$ for nodes of (out-) degree at least 3. We have $e = d = 3n - 3\ell + 3$, $s = n + 1 - \ell$, and $m = u = n + 1$, giving

$$\frac{1}{n+1}\binom{n+1}{\ell}\binom{2\ell - n - 3}{n - \ell}.$$

Summing over ℓ, for $2n/3 < \ell \leq n$, counts all n-edge trees sans unary and binary nodes. For $n = 3, 4, \ldots$, this is: 1, 1, 1, 4, 8, 13, 31, 71, 144, 318, 729, 1611, 3604, 8249, ... Note that this also counts the number of sequences of n natural numbers, excluding 1 and 2, such that the sum of every prefix is no more than its length.[3]

If we exclude only unary nodes, we get, instead, the nth Riordan number [21, A005043]. See [15, p. 587] and [7, Ex. 3.1.3]; see also [1]. □

Example 3.2. The number of ordered trees with n edges, r leaves, and i unary nodes is obtained by considering r leaf patterns $\Diamond$, i instances of $\langle \triangle \rangle$, and $n + 1 - r - i$ of $\langle \triangle\triangle\circ\circ \rangle$ for the remaining nodes ($e = d = 2n - 2r - i + 2$, $s = n - r - i + 1$, and $m = u = n + 1$):

$$\frac{1}{n+1}\binom{n+1}{r, i, n-r-i+1}\binom{r-2}{n-r-i}. \qquad \square$$

When patterns include all the nodes and all the edges in the tree ($e = n$, $m = u = d + 1$, and $s = 0$), enumeration (2.1) simplifies to just

$$\frac{1}{m}\binom{m}{n_1, \ldots, n_q}. \tag{3.1}$$

This generalizes [11] and [24], which consider patterns representing the degrees of the nodes only.

Example 3.3. The number of "0-1-2" ("unary-binary") trees (with maximum outdegree 2) with n edges and ℓ leaves, and, hence, $\ell - 1$ binary nodes and $n - 2\ell + 2$ unary nodes, is $\frac{1}{n+1}\binom{n+1}{\ell, \ell-1, n-2\ell+2}$ ([9]; cf. [19]). Summing over ℓ, we get

$$M_n = \frac{1}{n+1}\sum_{\ell}\binom{n+1}{\ell, \ell-1, n-2\ell+2}$$

for the total number of 0-1-2 trees of size n, which is the nth Motzkin number [21, A001006]. (A different derivation for this enumeration is given in [7, Ex. 3.1.1].) □

[3] This sequence was recently added as #A114997 to Neil Sloane's *The on-line encyclopedia of integer sequences* [21].

If one is interested only in a (lone) pattern occurring at the root of a tree (with no sibling ellipses, so the pattern does not overlap itself), then we have the following tree enumeration (see [7, sect. 3.3]):

$$\frac{d+s}{n-e}\binom{2n+d+s-2e-1}{n-e-1}. \tag{3.2}$$

When $n=e$, this is taken to be 1.

4. Patterns with nonleaf slots. Now we extend (2.1) by incorporating nonleaf patterns ▲. Like a plain triangle △, a dark triangle ▲ may also overlap an occurrence of another pattern (at its root), but with the added proviso that the latter is not ◊. This is why we need dark triangles, and cannot simply use ⟨△∞⟩ for nonleaf nodes, instead (something that can be done in the absence of leaf patterns).

For example, the pattern ⟨◊△∞▲⟩ (see Figure 2(c)) occurs in a tree, such as that in Figure 1(a), at each node of degree at least three, with eldest child a leaf and youngest not a leaf. The pair of patterns {⟨◊△∞▲⟩, ◊}, which includes another leaf pattern, occurs three times in the tree ⟨◊ ◊ ⟨◊ ◊⟩⟩ of Figure 1(a), once for each of the leaves, excluding the eldest child of the root. It occurs a total of 12 times in the 42 five-edged trees, but only in the following five of them:

$$\langle\Diamond\,\bar{\Diamond}\,\bar{\Diamond}\,\langle\bar{\Diamond}\rangle\rangle \quad \langle\Diamond\,\bar{\Diamond}\,\langle\langle\bar{\Diamond}\rangle\rangle\rangle \quad \langle\Diamond\,\bar{\Diamond}\,\langle\bar{\Diamond}\,\bar{\Diamond}\rangle\rangle \quad \langle\Diamond\,\langle\bar{\Diamond}\rangle\,\langle\bar{\Diamond}\rangle\rangle \quad \langle\langle\Diamond\,\bar{\Diamond}\,\langle\bar{\Diamond}\rangle\rangle\rangle.$$

The composite pattern matches at the high-degree node and the leaf pattern ◊ matches *any* one of the overlined leaves $\bar{\Diamond}$.

THEOREM 4.1. *Let $p_1,\dots,p_q$, $q\ge 0$, be various nonleaf nonellipsis patterns. The number of occurrences among all n-edge ordered trees of n_i of each of the patterns p_i and of $\ell\ge 0$ leaf patterns ◊ is*

$$\binom{m}{n_1,\dots,n_q}\binom{u-t}{\ell}\sum_{\ell\le k\le u}\frac{1}{u-k+1}\binom{u-k+1}{m}\binom{u-t-\ell}{k-\ell}\binom{n+s-e-1}{u+s-m-k}, \tag{4.1}$$

where $e=\Sigma e(p_i)$ is the total number of edges in the patterns, d is the number of plain triangles △ appearing in them, t is the number of dark triangles ▲, s is the number of ellipses ∞, $u=n+d+t-e$, and $m=\Sigma n_i$. When $m>0$, this enumeration can be rewritten as

$$\frac{1}{m}\binom{m}{n_1,\dots,n_q}\binom{u-t}{\ell}\sum_{\ell\le k\le u}\binom{u-k}{m-1}\binom{u-t-\ell}{k-\ell}\binom{n+s-e-1}{u+s-m-k}. \tag{4.1a}$$

Proof. We count separately for each possible number of "loose" (unattached to composite patterns) tree leaves, $k=\ell,\ell+1,\dots,u$. Note that the total number of leaves missing from the m patterns cannot exceed $n-e+d\le u$.

Let $v=\Sigma v(p_i)+\ell=m+e-d-t+k$ be the number of tree nodes accounted for by the patterns and leaves. (The $m+k$ patterns contain e edges, so the number of nodes and triangles is $v+d+t=e+m+k$.) The proof proceeds in several steps:

1. Arrange the given m nonleaf patterns in a row, in any of $\binom{m}{n_1,\dots,n_q}$ ways.
2. Intersperse $n+1-v=u+1-m-k$ extra patterns of the form ⟨△∞⟩ among the m patterns, to cover all the missing internal (nonleaf) nodes, in $\binom{u-k+1}{m}$ ways, for a total of $u-k+1$ patterns.

3. Distribute the $(n-e)-(n+1-v)=m-d-t+k-1$ missing edges (of the $n-e$ missing from the given patterns, $n+1-v$ were just added in the previous step), as sequences $\triangle\cdots\triangle$ of triangles, in place of the $s+n+1-v$ ellipses (s original and $n+1-v$ new), in $\binom{n+s-e-1}{m-d-t+k-1}=\binom{n+s-e-1}{u+s-m-k}$ ways. Note that when there are no missing edges ($n=e$ and $u+1-k=m$), this factor is $\binom{s-1}{0}=\binom{s-1}{s-1}=1$.
4. Place the ℓ distinguished leaves in some of the $d+(n+1-v)+(v-e-1)=u-t$ unrestricted, plain triangles (d in the original patterns, plus $n+1-v$ from step 2 and $v-e-1$ from step 3), in $\binom{u-t}{\ell}$ ways.
5. Place the remaining $k-\ell$ leaves in some of the remaining $u-t-\ell$ unrestricted triangles in $\binom{u-t-\ell}{k-\ell}$ ways.[4]
6. The cyclic arrangement of the resultant $m+(u+1-m-k)=u-k+1$ patterns corresponds to exactly one occurrence of the patterns in a tree. To see this, graft the patterns into one tree by repeatedly picking any pattern in the sequence and inserting it into the closest (rightmost) available triangle slot among the patterns preceding it, wrapping back around from the end when necessary. The $u-k+1$ patterns contain a total of $d+t+(u+1-m-k)+(m-d-t+k-1)-\ell-(k-\ell)=u-k$ slots. So, in fact, a single tree results from the grafting, with each pattern occurring at the point it ends up in the reconstructed tree. This situation may be viewed as an application of the cycle lemma, given in the introduction. Reading each pattern as the number of its slots, the lemma asserts that each of the $u-k+1$ cyclic permutations of the patterns gives one and the same grafted outcome. Thus, the enumeration has an additional factor of $\frac{1}{u-k+1}$.

Summing for k, and replacing $\frac{1}{u-k+1}\binom{u-k+1}{m}$ by $\frac{1}{m}\binom{u-k}{m-1}$ when $m>0$, gives the result. □

By way of illustration, let us count occurrences in 9-node trees of the quartet of patterns

written linearly as

$$\{\langle\circ\circ\circ\,\blacktriangle\rangle,\ \langle\circ\circ\circ\,\blacktriangle\rangle,\ \langle\triangle\Diamond\circ\circ\circ\,\rangle,\ \Diamond\}.$$

The patterns account for $e=4$ out of $n=8$ edges and for $v=5$ of the nodes, including 2 leaves, $\ell=1$ of which is a pattern on its own. There are $d=1$ plain triangles and $t=2$ dark triangles in the patterns, so $u=8+1+2-4=7$. The patterns also have $s=3$ ellipses.

The construction proceeds as follows:

0. Consider $k=3$, meaning that we want 2 more leaves in the tree, besides the lone $\Diamond$ appearing in the pattern set—and besides the one embedded in the third pattern, for a total of 3 internal nodes and 4 leaves.

[4]Steps 4 and 5 can be supplanted by first, (4′) choosing all k leaves in $\binom{n+d-e}{k}$ ways, and only then, (5′) selecting $\binom{k}{\ell}$ of them, for the same total contribution of $\binom{n+d-e}{k}\binom{k}{\ell}=\binom{n+d-e}{\ell}\binom{n+d-e-\ell}{k-\ell}$.

1. The $m = 3$ nonleaf patterns can be arranged in $\binom{3}{2,1} = 3$ ways, one of which is

$$\langle \circ\circ\circ \blacktriangle \rangle\,,\ \langle \triangle \Diamond \circ\circ\circ \rangle\,,\ \langle \circ\circ\circ \blacktriangle \rangle\,.$$

2. We need to add 2 internal-node patterns $\langle \triangle\circ\circ\circ \rangle$ for the remaining 2 as yet unaccounted-for nodes. These extra patterns can be intermingled with the 3 original nonleaf patterns in $\binom{3+2}{3} = 10$ different ways, including, for example, the following sequence:

$$\langle \triangle\circ\circ\circ \rangle\,,\ \langle \circ\circ\circ \blacktriangle \rangle\,,\ \langle \triangle\circ\circ\circ \rangle\,,\ \langle \triangle \Diamond \circ\circ\circ \rangle\,,\ \langle \circ\circ\circ \blacktriangle \rangle\,.$$

3. There are only 6 edges in these patterns, so we need to graft in the 2 remaining tree edges, and eliminate all $3 + 2 = 5$ ellipses in the process. This can be done in $\binom{5+2-1}{2} = 15$ ways, such as

$$\langle \triangle \rangle\,,\ \langle \triangle \blacktriangle \rangle\,,\ \langle \triangle \rangle\,,\ \langle \triangle \Diamond \triangle \rangle\,,\ \langle \blacktriangle \rangle\,.$$

4. The $\ell = 1$ leaf in the original pattern set can go into any of the 5 ordinary subtree ($\triangle$) slots—in the middle one, say,

$$\langle \triangle \rangle\,,\ \langle \triangle \blacktriangle \rangle\,,\ \langle \Diamond \rangle\,,\ \langle \triangle \Diamond \triangle \rangle\,,\ \langle \blacktriangle \rangle\,.$$

5. Similarly, the extra $k - \ell = 2$ leaves can go into any of the remaining $n + d - e - \ell = 4$ plain slots, in $\binom{4}{2} = 6$ combinations, giving something such as this final set of 5 tree patterns, arranged in sequence:

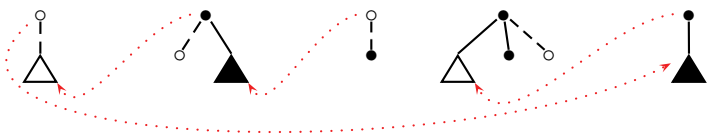
$$\langle \triangle \rangle\,,\ \langle \Diamond \blacktriangle \rangle\,,\ \langle \Diamond \rangle\,,\ \langle \triangle \Diamond \Diamond \rangle\,,\ \langle \blacktriangle \rangle\,.$$

   We have 4 slots now and 5 patterns. All the leaves have been allocated, so the type of slot no longer matters.

6. Last, these 5 tree pieces are grafted together, step by step, as follows:

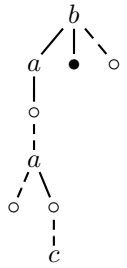
$$\begin{gathered}
\langle \triangle \rangle\,,\ \langle \Diamond \blacktriangle \rangle\,,\ \langle \Diamond \rangle\,,\ \langle \triangle \Diamond \Diamond \rangle\,,\ \langle \blacktriangle \rangle \\
\langle \langle \Diamond \blacktriangle \rangle \rangle\,,\ \langle \Diamond \rangle\,,\ \langle \triangle \Diamond \Diamond \rangle\,,\ \langle \blacktriangle \rangle \\
\langle \langle \Diamond\, \langle \Diamond \rangle \rangle \rangle\,,\ \langle \triangle \Diamond \Diamond \rangle\,,\ \langle \blacktriangle \rangle \\
\langle \langle \Diamond\, \langle \Diamond \rangle \rangle \rangle\,,\ \langle \langle \blacktriangle \rangle\, \Diamond \Diamond \rangle \\
\langle \langle \langle \langle \Diamond\, \langle \Diamond \rangle \rangle \rangle \rangle\, \Diamond \Diamond \rangle\,.
\end{gathered}$$

   Pictorially, this is what happens:



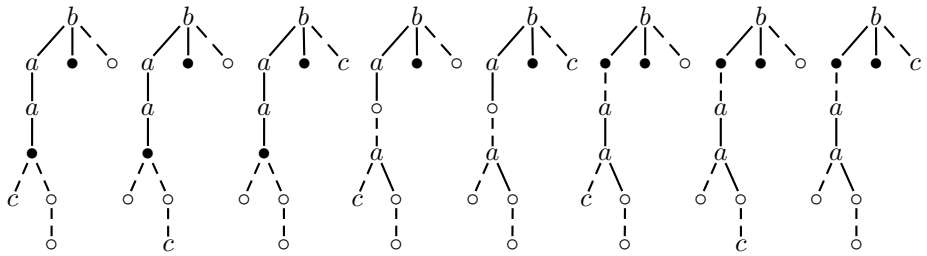
   A dashed line in a pattern means that the edge is not from the original pattern set and similarly for an unfilled node. One can see, then, that the original patterns account for 4 out of 8 edges, 2 out of 4 leaves, and 3 out of 5 internal nodes.

The point of the cycle lemma is that the outcome is the same regardless of the order in which (cyclic) neighbors are grafted. The same end result, namely,



would be obtained from any of the 5 cyclic permutations of the starting sequence. The third of the original four patterns, $\langle \triangle \Diamond \infty\infty \rangle$, occurs at the root, marked b; the first two patterns, $\langle \infty\infty \blacktriangle \rangle$, occur at the nodes marked a; the distinguished leaf is c.

The above theorem says that there are a total of $3 \times 10 \times 15 \times 5 \times 6 \times \frac{1}{5} = 2700$ occurrences of the given patterns among all 490 trees with 8 edges and 4 leaves. For example, the same tree as obtained above happens to contain 8 additional occurrences of the same multiset of patterns, as follows:



Example 4.2. As a trivial example, with no patterns, this formula counts ordered trees of size n, by summing the number of trees with $k = 1, \ldots, n+1$ leaves, since a tree has only one occurrence of an empty pattern set. Letting $q = m = e = d = s = t = \ell = 0$ and $u = n$ in (4.1), we obtain

$$\sum_k \frac{1}{n+1-k}\binom{n}{k}\binom{n-1}{k-1} = \frac{1}{n}\sum_k \binom{n}{k}\binom{n}{k-1} = \frac{1}{2n+1}\binom{2n+1}{n},$$

the nth Catalan number, C_n. Each term $\frac{1}{n}\binom{n}{k-1}\binom{n}{k}$ in the sum is a Narayana number [17], and represents the number of n-edge k-leaf trees, as shown in [6, 18]. □

Remark 4.3. Suppose there are no dark triangle ($\blacktriangle$) subpatterns ($t = 0$, $u = n + d - e$). Then (4.1) reduces to (2.1), but with an additional $n_{q+1} = \ell$ occurrences of the leaf pattern:

$$\binom{m}{n_1, \ldots, n_q}\binom{u-1}{\ell}\sum_k \frac{1}{u-k+1}\binom{u-k+1}{m}\binom{u-\ell}{k-\ell}\binom{n+s-e-1}{u+s-m-k}$$

$$= \sum_k \frac{1}{u-m-k+1}\binom{u}{n_1, \ldots, n_q, \ell, k-\ell, u-m-k}\binom{n+s-e-1}{u+s-m-k}$$

$$= \frac{1}{u+1}\binom{u+1}{n_1,\ldots,n_q,\ell,u-m-\ell+1}\sum_k \binom{u-m-\ell+1}{k-\ell}\binom{n+s-e-1}{u+s-m-k}$$

$$= \frac{1}{u+1}\binom{u+1}{n_1,\ldots,n_q,\ell,u-m-\ell+1}\binom{2n+d+s-2e-m-\ell}{n-e}.$$

When the patterns account for all the leaves (as, for instance, when they account for all $n+1$ nodes), the sum contributes only the $k=\ell$ term, and (4.1a) simplifies to

$$\frac{1}{m}\binom{u-\ell+1}{n_1,\ldots,n_q,u-\ell-m+1}\binom{n+d-e}{\ell}\binom{n+s-e-1}{u+s-m-\ell}. \tag{4.2}$$

If they also account for all edges ($n=e$), then $u=d+t=m+k-1=m+\ell-1$, and we get the following mild extension of (3.1):

$$\frac{1}{m}\binom{m}{n_1,\ldots,n_q}\binom{d}{\ell}, \tag{4.3}$$

with an added factor for the $\binom{d}{\ell}$ ways of filling ℓ of the d unrestricted slots with leaves.

Example 4.4. The number of binary trees with i internal nodes and j left leaves is also counted by the Narayana numbers, as is clear from the standard correspondence [15, sect. 2.3.2] between ordered trees with i edges and binary trees with i internal nodes. Let there be $n_1=j$ patterns $\langle\Diamond\triangle\rangle$, $n_2=t=i-j$ patterns $\langle\blacktriangle\triangle\rangle$, and $\ell=i+1-j$ leaf patterns $\Diamond$. Then put $d=i$, $n=e=2i$, and $s=0$ (hence, $m=i$ and $u=2i-j$) into the formula. Only $k=\ell$ contributes to the sum, and the last factor in (4.2) is $\binom{-1}{0}=1$, so we get

$$\frac{1}{i}\binom{i}{j}\binom{i}{i+1-j} = \frac{1}{i}\binom{i}{j}\binom{i}{j-1}. \qquad \square$$

Example 4.5. The number of ordered trees with n edges, k leaves, and j leftmost (that is, "eldest") nonroot internal nodes (out of a total of $n-k$ nonroot internal nodes) is counted by $n_1=j$ patterns p_1 of the form $\langle\blacktriangle\infty\infty\rangle$, $n_2=n+1-j-k$ of the form $p_2=\langle\Diamond\infty\infty\rangle$ (for the internal nodes not covered by p_1), and $\ell=k-n_2=2k+j-n-1$ leaf patterns (for the leaves not in p_2). Putting $m=e=s=n-k+1$, $t=j$, $d=0$, and $u=j+k-1$ into (4.2), we have

$$\frac{1}{n-k+1}\binom{n-k+1}{j}\binom{k-1}{2k+j-n-1}\binom{n-1}{n-k}$$

$$= \frac{1}{n}\binom{n}{j,n-j-k+1,n-j-k,2k+j-n-1}.$$

Letting $i=n-j-k+1$ yields

$$\frac{1}{n}\binom{n}{i,i-1,j,n-2i-j+1}$$

for the number of trees with j leftmost internal nodes, not counting the root, and $i-1$ nonleftmost ones. The tree in Figure 1(a), for instance, has one nonleftmost internal node, but no leftmost ones (other than the root). $\square$

Example 4.6. Looking back at Example 3.2, suppose one wishes to count leftmost leaves separately. Let there be n edges, q internal nodes, i of which are unary, and

$r = n+1-q$ leaves, j of which are leftmost. We must, therefore, distinguish between nodes with leftmost leaves, and those without. If x is the number of unary nodes with a lone-leaf child, then we want x occurrences of $\langle\Diamond\rangle$, $i-x$ of $\langle\blacktriangle\rangle$, $j-x$ of $\langle\Diamond\triangle\infty\infty\rangle$, $\ell = r-j$ loose leaves $\Diamond$, and $n+1-i+x-2j-\ell = q+x-i-j$ of $\langle\blacktriangle\triangle\infty\infty\rangle$, making $m = n-r+1$, $e = 2n-2r-i+2$, $d = s = n-r-i+1$, $t = n-j-r+1$, and $u = n-j$. Summing over x, we get

$$\frac{1}{q}\binom{r-1}{r-j}\binom{r-2}{r+i-q-1}\sum_x \binom{q}{x, i-x, j-x, q+x-i-j}$$

$$= \frac{1}{q}\binom{r-1}{j-1}\binom{r-2}{q-i-1}\binom{q}{i}\sum_x \binom{i}{x}\binom{q-i}{j-x}$$

$$= \frac{1}{q}\binom{q}{i}\binom{q}{j}\binom{r-1}{j-1}\binom{r-2}{q-i-1}. \qquad \square$$

5. Up and down steps. By the standard correspondence between binary trees and Dyck (nonnegative lattice) paths (or *bridges*) [18], the number of paths from $(0,0)$ to (m,m) not crossing below the baseline (grid diagonal), with exactly j occurrences of the lattice pattern **udu** (a step up, **u**, followed by a step down, **d**, and another step up), as counted in [23], is the same as the number of occurrences of binary trees that have j left leaves with a nonleaf sibling amongst binary trees with m internal nodes. For example, the lattice path in Figure 1(c) has 3 occurrences of **udu**, drawn sideways as $\nearrow\!\searrow\!\nearrow$, corresponding to the first (reading left to right) three left leaves in Figure 1(b).

We can count these tree patterns by using (3.1), taking $n_1 = j$ patterns $\langle\Diamond\triangle\rangle$ for internal nodes with a left leaf, $n_2 = i$ of $\langle\Diamond\Diamond\rangle$ for "double" leaves (internal nodes with two leaves), $n_3 = m+1-j-2i$ of $\langle\triangle\Diamond\rangle$ for internal nodes with a right leaf, and $n_4 = i-1$ for the rest $\langle\triangle\triangle\rangle$, for a total of m patterns ($e = 2m$, $d = m-1$) and

$$\frac{1}{m}\binom{m}{j, i, m-2i-j+1, i-1} \tag{5.1}$$

occurrences. Since the patterns account for all the internal nodes of the tree, no triangle subpattern $\triangle$ can be filled by anything but another binary node.

Summing over all i, we get

$$\frac{1}{m}\binom{m}{j}\sum_i \binom{m-j}{i, i-1, m-2i-j+1}$$

$$= \frac{1}{m}\binom{m}{j}\sum_i \binom{m-j}{i}\binom{m-i-j}{i-1}$$

for the number of binary trees with m binary nodes, j of which have a left leaf child and right nonleaf. For $m = 5$ and $j = 3$, there are $\frac{1}{5}\binom{5}{3}\binom{2}{1} = 4$ such binary trees, including the one portrayed in Figure 1(b).

This enumeration is equivalent to the formula

$$\binom{m-1}{j}M_{m-j-1} = \frac{1}{m-j}\binom{m-1}{j}\sum_i \binom{m-j}{i, i-1, m-2i-j+1}$$

in [23, Thm. 2.1] for Dyck paths with m **u**-steps and m **d**-steps, and including j segments **udu**, where M_{m-j-1} is a Motzkin number (as in Example 3.3).

Alternatively, we can apply (4.1a) with $n_1 = j$ patterns $\langle \Diamond \blacktriangle \rangle$ for internal nodes with a leftmost leaf and rightmost nonleaf, $n_2 = i$ double-leaf patterns $\langle \Diamond \Diamond \rangle$, and $n_3 = m - j - i$ of $\langle \blacktriangle \triangle \rangle$ for the remaining internal nodes, for a total of m patterns ($n = e = 2m$, $d = m - j - i$, $t = m - i$, $s = \ell = 0$, $u = 2m - 2i - j$). Recalling the convention that $\binom{-1}{0} = 1$, we have

$$\begin{aligned}
&\frac{1}{m}\binom{m}{j, i, m-j-i}\sum_k \binom{m-j-i}{k}\binom{-1}{m-u+k+1} \\
&= \frac{1}{m}\binom{m}{j, i, m-j-i}\binom{m-j-i}{u-m+1} \\
&= \frac{1}{m}\binom{m}{i, i-1, j, m-2i-j+1},
\end{aligned}$$

as before.

A *low-level* occurrence of a lattice pattern, like **udu**, means that one end of each **u** and **d** step touches the baseline. For example, the lattice path in Figure 1(c) has two low occurrences of **udu** (the first two such occurrences). To count these, we use, this time, the standard correspondence between paths and ordered trees [22], in which the two low occurrences in Figure 1(c) correspond to the two first leaves in Figure 1(a).

For paths from $(0, 0)$ to (n, n) with j low occurrences, we need to count n-edge trees with exactly j nonyoungest (i.e., nonrightmost) leaves sprouting from the root. For example, the pattern $\langle \langle \triangle \infty \infty \rangle \langle \triangle \infty \infty \rangle \Diamond \triangle \rangle$ matches a root of degree 4, with the third child having no children, but its two older siblings having children. For a root of degree r, there are, in general, $\binom{r-1}{j}$ such patterns for the different placements of the j leaves (in every place but the last).

Substituting $s = r - j - 1$, $d = s + 1$, and $e = r + s$ into our root formula (3.2), and summing over root degree r, we get

$$\sum_r \frac{2r-2j-1}{n+j-2r+1}\binom{r-1}{j}\binom{2n-2j}{n-r-1}$$

(ignoring the fraction whenever the denominator is 0) for the number of paths with j low-level **udu**'s, as counted in [23, Thm. 3.1]. For example, for $n = 5$ and $j = 2$, we get $\frac{1}{2}\binom{6}{1}\binom{2}{2} + \binom{6}{0}\binom{3}{2} = 6$, including the path in Figure 1(c).

6. Young and old leaves. The number of ordered trees with n edges, i oldest (leftmost) leaves, and j younger (nonleftmost) leaves is given by (4.2), with $n_1 = i$ nodes $\langle \Diamond \infty \infty \rangle$ with eldest leaf, $n_2 = t = n + 1 - 2i - j$ nodes $\langle \blacktriangle \infty \infty \rangle$ with nonleaf eldest, $m = e = s = n - i - j + 1$, $d = 0$, and $u = n - i$ (all leaves are accounted for):

$$\begin{aligned}
&\frac{1}{n-i-j+1}\binom{n-i-j+1}{i}\binom{i+j-1}{j}\binom{n-1}{n-i-j} \\
&= \frac{1}{n}\binom{n}{i, i-1, j, n-2i-j+1} \qquad (6.1) \\
&= \frac{1}{n}\binom{n}{j}\binom{n-j}{i}\binom{n-i-j}{i-1}.
\end{aligned}$$

The above formula counts trees, since all nodes are covered by the m patterns and j leaves. It is equivalent to the enumeration in [4, Prop. 2.1], namely,

$$\frac{1}{n}\binom{n}{i}\binom{n-i}{j}\binom{n-i-j}{i-1}.$$

There, a tree-grafting proof, based on [3], and similar to the idea in [7] that has been reused here, is provided. The tree in Figure 1(a) is one of the $\frac{1}{5}\binom{5}{2,1,2} = 6$ five-edge trees with 2 oldest leaves and 2 younger ones.

Summing the formula for all i gives the total number of n-edge trees containing j noneldest leaves:

$$\frac{1}{n}\binom{n}{j}\sum_i \binom{n-j}{i}\binom{n-i-j}{i-1}.$$

By Proposition 2.2, a tree with i leftmost leaves and j nonleftmost ones has $i-1$ nonleftmost interior nodes and $n-2i-j+2$ leftmost ones. Letting $r = n-2i-j+2$ and $s = i-1$, we have that there are

$$\frac{1}{n}\binom{n}{r-1, s, s+1, n-2s-r}$$

trees with n edges, r leftmost interior nodes, and s nonleftmost ones, and that there are a total of

$$\frac{1}{n}\binom{n}{r-1}\sum_s \binom{n-r+1}{s, s+1, n-2s-r}$$

n-edge trees with r leftmost internal nodes and any number of nonleftmost ones.

7. Odd and even levels. The Narayana numbers,

$$\frac{1}{n}\binom{n}{q}\binom{n}{q-1},$$

also happen to count the number of ("bicolored") n-edge ordered trees with q nodes (leaves or internal) on their odd levels (the root's children, great-grandchildren, etc.) [24]. (See [5, Thm. 4.3B].) For example, $q = 3$ in Figure 1(a). This enumeration is refined in [5, Thm. 4.4B], where it is shown that there are

$$\frac{1}{r}\binom{r}{\ell}\binom{q}{i}\binom{r-2}{q-i-1}\binom{q-1}{r-\ell-1}$$

trees with q odd nodes, including i leaves, and $r = n+1-q$ even nodes, ℓ of which are leaves. For example, $i = \ell = 2$ for the tree in Figure 1(a).

In [6], we gave the following bijection between n-edge ordered trees with ℓ leaves and those with ℓ internal nodes:

$$\langle\rangle^\sharp = \langle\rangle,$$

$$\langle s, t_1, \ldots, t_k\rangle^\sharp = \langle\langle t_1, \ldots, t_k\rangle^\sharp, s_1, \ldots, s_m\rangle,$$

where

$$s^\sharp = \langle s_1, \ldots, s_m\rangle.$$

Applying that idea to even levels, but leaving odd levels intact, we get a bijection defined as follows:

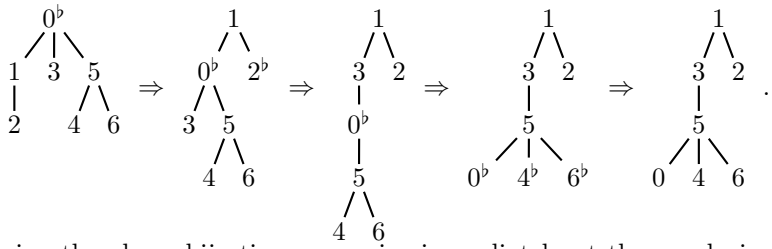
$$\begin{aligned} \langle\rangle^\flat &= \langle\rangle, \\ \langle\langle s_1,\ldots,s_m\rangle,t_1,\ldots,t_k\rangle^\flat &= \langle\langle t_1,\ldots,t_k\rangle^\flat, s_1^\flat,\ldots,s_m^\flat\rangle. \end{aligned} \tag{7.1}$$

This maps n-edge trees ($n>0$) with j even internal nodes, ℓ even leaves, and i odd leaves to n-edge trees with j leftmost leaves, ℓ nonleftmost leaves, and i unary nodes. More generally, odd nodes are mapped via this bijection to internal nodes, each with one additional child.

To see the correspondences, consider the following points:

- The above mapping $\flat$ is applied only to internal nodes from even levels; the recursion continues down the leftmost branch until only a leaf remains. So each even internal node corresponds to a leftmost leaf.
- If one of the s_i is a leaf, then it maps to $s_i^\flat = \langle\rangle^\flat = \langle\rangle$, yielding a nonleftmost leaf.
- Each odd node $\langle s_1,\ldots,s_m\rangle$ of degree m corresponds to the degree $m+1$ internal node that results from $\langle\langle t_1,\ldots,t_k\rangle^\flat, s_1^\flat,\ldots,s_m^\flat\rangle$.

The following example serves to illustrate the process:



Applying the above bijection, we arrive immediately at the conclusion that the formula of Example 4.6, viz.,

$$\begin{aligned} &\frac{1}{q}\binom{q}{i}\binom{q}{r-\ell}\binom{r-1}{\ell}\binom{r-2}{q-i-1} \\ &= \frac{1}{r}\binom{r}{\ell}\binom{q}{i}\binom{r-2}{q-i-1}\binom{q-1}{r-\ell-1} \\ &= \frac{1}{r}\binom{r}{j}\binom{q}{k}\binom{r-2}{k-1}\binom{q-1}{j-1}, \end{aligned}$$

also counts trees with q odd nodes, r even nodes, i odd leaves, j even internal nodes, $k=q-i$ odd internal nodes, and $\ell = r-j$ even leaves, rederiving the enumeration in [5].

There is also a simple bijection between trees that flips odd and even levels, for trees with at least two levels, and preserves the degree of all but two nodes:

$$\langle\rangle^\circ = \langle\rangle,$$

$$\langle\langle s_1,\ldots,s_m\rangle,t_1,\ldots,t_k\rangle^\circ = \langle\langle t_1,\ldots,t_k\rangle,s_1,\ldots,s_m\rangle.$$

We have the following theorem.

THEOREM 7.1.

(1) *The expected degree of even-level nodes among all ordered trees of a given size is exactly* 1.

(2) *The expected degree of odd-level nodes among all size n ordered trees is* $\frac{n-1}{n+1}$.

Proof. The expected degree of an even node is exactly 1, since

$$\text{average even degree} = \frac{\text{total number of odd nodes}}{\text{total number of even nodes}} = 1.$$

The first equality is by definition; the second is on account of the odd to even bijection.

Since there are the same quantities of odd and even nodes, we must have

$$\text{average odd degree} = 2 \times \text{average degree} - \text{average even degree} = 2\frac{n}{n+1} - 1.$$

The expected degree of an arbitrary node is clearly $\frac{n}{n+1}$, there being $n+1$ nodes in each tree of size n. □

To summarize, taking Proposition 2.2 into account, we have the following theorem.

THEOREM 7.2.

(1) *The following distributions in ordered trees of a given size (greater than* 0*) are identical:*
   (a) *even-level nodes (per tree);*
   (b) *odd-level nodes;*
   (c) *leaves;*
   (d) *internal nodes.*

(2) *The following distributions in ordered trees of a given size (greater than* 0*) are identical:*
   (a) *even-level leaves;*
   (b) *younger leaves;*
   (c) *eldest internal nodes, minus* 1.

(3) *The following distributions in ordered trees of a given size (greater than* 0*) are identical:*
   (a) *even-level internal nodes;*
   (b) *eldest leaves;*
   (c) *younger internal nodes, plus* 1.

(4) *The following distributions in ordered trees of a given size are identical:*
   (a) *odd-level nodes of degree d;*
   (b) *nodes of degree $d+1$.*

Thus, statistics for the distribution of node degrees can be applied to the degrees of odd nodes—with an offset of 1. For example, since the average degree of an internal node is $2n/(n+1) \approx 2$ [6, Cor. 2.2], the average degree of all odd nodes (leaf or internal) is $2n/(n+1) - 1 = (n-1)/(n+1) \approx 1$.

The above correspondences explain why, in fact, the Narayana numbers count trees with a given number of odd (even) nodes, just as they do trees with a given number of internal nodes (leaves). See Example 4.2.

Since the bijection between odd-level nodes and even-level nodes changes the degree of at most two nodes and can turn at most one leaf into an internal node, we also have the following theorem.

THEOREM 7.3.

(1) *The following distributions in ordered trees of a given size (greater than* 0*) are identical, give or take* 1*:*

(a) *even-level leaves, or younger leaves (per tree);*
(b) *even-level internal nodes, or eldest leaves;*
(c) *odd-level leaves;*
(d) *odd-level internal nodes;*
(e) *eldest internal nodes;*
(f) *younger internal nodes.*

(2) *The following distributions in ordered trees of a given size (greater than* 0*) are identical, give or take* 1*:*
(a) *odd-level nodes of degree d;*
(b) *even-level nodes of degree d.*

We know that leaves and internal nodes have the same distributions and that odd and even nodes also do. Furthermore, the cases in which the odd-even bijection changes an odd-level leaf into an even-level internal node are precisely those when the leftmost level-one node is a leaf. There are exactly C_{n-1} such cases for trees of size n. Accordingly, we have the following theorem.

THEOREM 7.4.

(1) *The expected number of even-level leaves in a size n ordered tree is*

$$\frac{n+1}{4} - \frac{1}{2}\frac{C_{n-1}}{C_n} = \frac{n+1}{4} - \frac{n+1}{4(2n-1)} = \frac{n+1}{4}\left(1 - \frac{1}{2n-1}\right).$$

(2) *The expected number of odd-level leaves in a size n ordered tree is*

$$\frac{n+1}{4} + \frac{1}{2}\frac{C_{n-1}}{C_n} = \frac{n+1}{4}\left(1 + \frac{1}{2n-1}\right).$$

(3) *The expected number of even-level internal nodes in a size n ordered tree is*

$$\frac{n+1}{4}\left(1 + \frac{1}{2n-1}\right).$$

(4) *The expected number of odd-level internal nodes in a size n ordered tree is*

$$\frac{n+1}{4}\left(1 - \frac{1}{2n-1}\right).$$

8. Up-down-ups, younger leaves, and even leaves. It should come as no surprise at this point that the same number, namely,

$$(\star) \qquad \frac{1}{n}\binom{n}{i, i-1, j, n-2i-j+1},$$

counts

a. ordered trees with n edges, i oldest leaves, and j younger ones (see (6.1) and [4]);
b. binary trees with n internal nodes, i internal nodes with two leaves, and j internal nodes with a leaf only on the left (see (5.1));
c. Dyck (nonnegative lattice) paths with n **u**'s, $n+1$ **d**'s, i **udd**'s, and j **udu**'s, where the extra **d** is always placed at the end of the path and cannot affect the **udu** count (cf. section 5 and [23]);
d. ordered trees with n edges, i even internal nodes, and j even leaves (cf. section 7 and [5]); as well as

e. ordered trees with n edges, $j+1$ eldest internal nodes, and $i-1$ younger internal nodes (Example 4.5).

For the correspondences between enumerations (a), (b), (c), just consider the standard correspondences [15, 6, 22], under which a leftmost leaf (or, symmetrically, a rightmost leaf) in an ordered tree, a double leaf in a binary tree, and a lattice-path sequence **udd** all correspond to each other, as do a nonleftmost (or nonrightmost) leaf in an ordered tree, a lone leftmost leaf in a binary tree, and a path sequence **udu**. For the correspondence between (a) and (d), use bijection $\flat$ (see (7.1)), matching even internal nodes with oldest leaves and even leaves with younger leaves. For the correspondence between (a) and (e), recall (from Proposition 2.2) that the number of eldest leaves in a tree equals the number of younger internal nodes plus the root, and (from Theorem 7.2(2)) that younger leaves match up with eldest internal nodes minus the root.

Rewriting formula ($\star$) in terms of a total of $\ell = i + j$ leaves in an n-edge ordered tree, i of which are leftmost, we get

$$\frac{1}{n}\binom{n}{i}\binom{n-\ell}{i-1}\binom{n-i}{n-\ell}.$$

Summing over all i gives the Narayana numbers,

$$\frac{1}{n}\binom{n}{\ell}\binom{n}{\ell-1},$$

for trees with n edges and ℓ leaves (leftmost or otherwise), as well as for trees with ℓ even nodes [24]. This, in turn (see Example 4.2), adds up to the Catalan number,

$$C_n = \frac{1}{2n+1}\binom{2n+1}{n},$$

for size n ordered trees with any number of leaves.

Acknowledgment. We thank the referees for their suggested improvements.

REFERENCES

[1] F. R. Bernhart, *Catalan, Motzkin, and Riordan numbers*, Discrete Math., 204 (1999), pp. 73–112.
[2] E. C. Catalan, *Note sur un problème de combinaisons*, J. Math. Pures Appl., 3 (1838), pp. 111–112.
[3] W. Y. C. Chen, *A general bijection algorithm for trees*, Proc. Natl. Acad. Sci. USA, 87 (1990), pp. 9635–9639.
[4] W. Y. C. Chen, E. Deutsch, and S. Elizalde, *Old and young leaves on plane trees*, European J. Combin., 27 (2006), pp. 414–427; available online from http://www.math.dartmouth.edu/∼sergi/papers/oldleaves.pdf.
[5] L. H. Clark, J. E. McCanna, and L. A. Székely, *A survey of counting bicoloured trees*, Bull. Inst. Combin. Appl., 21 (1997), pp. 33–45.
[6] N. Dershowitz and S. Zaks, *Enumerations of ordered trees*, Discrete Math., 31 (1980), pp. 9–28.
[7] N. Dershowitz and S. Zaks, *Patterns in trees*, Discrete Appl. Math., 25 (1989), pp. 241–255.
[8] N. Dershowitz and S. Zaks, *The cycle lemma and some applications*, European J. Combin., 11 (1990), pp. 35–40.
[9] R. Donaghey and L. W. Shapiro, *Motzkin numbers*, J. Combin. Theory Ser. A, 23 (1977), pp. 291–301.
[10] A. Dvoretsky and T. Motzkin, *A problem of arrangements*, Duke Math. J., 14 (1947), pp. 305–313.

[11] A. Erdélyi and I. M. H. Etherington, *Some problems of non-associative combinations* II, Edinburgh Math. Notes, 32 (1941), pp. 7–12.
[12] P. Flajolet and J.-M. Steyaert, *On the analysis of tree-matching algorithms*, in Proceedings of the 7th Colloquium on Automata, Languages and Programming, J. W. Bakker and J. van Leeuwen, eds., Lecture Notes in Comput. Sci. 85, Springer-Verlag, Berlin, 1980, pp. 208–219.
[13] H. W. Gould, *Catalan and Bell Numbers: Research Bibliography of Two Special Number Sequences*, 6th ed., Mathematica Monongaliae 12, Combinatorial Research Institute, Morgantown, WV, 1985.
[14] F. Harary, G. Prins, and W. T. Tutte, *The number of plane trees*, Indag. Math., 26 (1964), pp. 319–329.
[15] D. E. Knuth, *The Art of Computer Programming, Vol.* 1: *Fundamental Algorithms*, Addison-Wesley, Reading, MA, 1968.
[16] S. Gopal Mohanty, *Lattice Path Counting and Applications*, Academic Press, New York, 1979.
[17] T. Venkata Narayana, *Sur les treillis formés par les partitions d'un entier et leurs applications à la théorie des probabilités*, C. R. Acad. Sci. Paris, 240 (1955), pp. 1188–1189.
[18] J. Riordan, *Enumeration of plane trees by branches and endpoints*, J. Combin. Theory Ser. A, 19 (1975), pp. 215–222.
[19] G. Rote, *Binary trees having a given number of nodes with* 0*,* 1*, and* 2 *children*, Sém. Lothar. Combin., 38 (1996); available online from http://www.mat.univie.ac.at/∼slc/wpapers/s38pr_rote.pdf, 6 pp.
[20] J. A. von Segner, *Enumeratio modorum, quibus figurae planae rectilineae per diagonales dividuntur in triangula*, Novi Comm. Acad. Scient. Imper. Petropolitanae, 7 (1759), pp. 203–209.
[21] N. J. A. Sloane, *The on-line encyclopedia of integer sequences*, http://www.research.att.com/∼njas/sequences (2006).
[22] R. P. Stanley, *Enumerative Combinatorics, Vol.* I, Wadsworth & Brooks/Cole, Monterey, CA, 1986.
[23] Y. Sun, *The statistic "number of udu's" in Dyck paths*, Discrete Math., 287 (2004), pp. 177–186.
[24] W. T. Tutte and F. Harary, *The number of plane trees with a given partition*, Mathematika, 11 (1964), pp. 99–101.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 466–476

SPANNING DIRECTED TREES WITH MANY LEAVES*

NOGA ALON†, FEDOR V. FOMIN‡, GREGORY GUTIN§, MICHAEL KRIVELEVICH†, AND SAKET SAURABH‡

Abstract. The DIRECTED MAXIMUM LEAF OUT-BRANCHING problem is to find an out-branching (i.e., a rooted oriented spanning tree) in a given digraph with the maximum number of leaves. In this paper, we obtain two combinatorial results on the number of leaves in out-branchings. We show that (1) every strongly connected n-vertex digraph D with minimum in-degree at least 3 has an out-branching with at least $(n/4)^{1/3} - 1$ leaves; (2) if a strongly connected digraph D does not contain an out-branching with k leaves, then the pathwidth of its underlying graph $\mathrm{UG}(D)$ is $O(k \log k)$, and if the digraph is acyclic with a single vertex of in-degree zero, then the pathwidth is at most $4k$. The last result implies that it can be decided in time $2^{O(k \log^2 k)} \cdot n^{O(1)}$ whether a strongly connected digraph on n vertices has an out-branching with at least k leaves. On acyclic digraphs the running time of our algorithm is $2^{O(k \log k)} \cdot n^{O(1)}$.

Key words. out-branching, maximum leaf, fixed parameter tractability, rooted tree, directed graphs

AMS subject classifications. 05C05, 05C85, 68R10, 68W05

DOI. 10.1137/070710494

1. Introduction. In this paper, we initiate the combinatorial and algorithmic study of a natural generalization of the well-studied MAXIMUM LEAF SPANNING TREE (MLST) problem on connected undirected graphs [9, 14, 17, 21, 18, 22, 24, 31, 33]. Given a digraph D, a subdigraph T of D is an *out-tree* if T is an oriented tree with only one vertex s of in-degree zero (called *the root*). If T is a spanning out-tree, i.e., $V(T) = V(D)$, then T is called an *out-branching* of D. The vertices of T of out-degree zero are called *leaves*. The DIRECTED MAXIMUM LEAF OUT-BRANCHING (DMLOB) problem is to find an out-branching in a given digraph with the maximum number of leaves.

It is well known that MLST is NP-hard for undirected graphs [23], which means that DMLOB is NP-hard for symmetric digraphs (i.e., digraphs in which the existence of an arc xy implies the existence of the arc yx) and, thus, for strongly connected digraphs. We can show that DMLOB is NP-hard for acyclic digraphs as follows: Consider a bipartite graph G with bipartition X, Y and a vertex $s \notin V(G)$. To obtain an acyclic digraph D from G and s, orient the edges of G from X to Y and add all arcs sx, $x \in X$. Let B be an out-branching in D. Then the set of leaves of B is $Y \cup X'$, where $X' \subset X$, and for each $y \in Y$ there is a vertex $z \in Z = X \setminus X'$ such that $zy \in A(D)$. Observe that B has maximum number of leaves if and only if $Z \subseteq X$ is of

*Received by the editors December 10, 2007; accepted for publication (in revised form) September 17, 2008; published electronically January 16, 2009. Preliminary extended abstracts of this paper have been presented at FSTTCS 2007 [3] and ICALP 2007 [2].

http://www.siam.org/journals/sidma/23-1/71049.html

†Department of Mathematics, Tel Aviv University, Tel Aviv 69978, Israel (nogaa@post.tau.ac.il, krivelev@post.tau.ac.il). The research of these authors was supported in part by USA-Israeli BSF grants and by grants from the Israel Science Foundation.

‡Department of Informatics, University of Bergen, POB 7803, 5020 Bergen, Norway (fomin@ii.uib.no, saket@ii.uib.no). The research of the second author was supported in part by the Norwegian Research Council.

§Corresponding author. Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK (gutin@cs.rhul.ac.uk). This author's research was supported in part by EPSRC.

minimum size among all sets $Z' \subseteq X$ such that $N_G(Z') = X$. However, the problem of finding Z' of minimum size such that $N_G(Z') = X$ is equivalent to the Set Cover problem ($\{N_G(y)|\ y \in Y\}$ is the family of sets to cover), which is NP-hard.

The combinatorial study of spanning trees with maximum number of leaves in undirected graphs has an extensive history. Linial conjectured around 1987 that every connected graph on n vertices with minimum vertex degree δ has a spanning tree with at least $n(\delta-2)/(\delta+1)+c_\delta$ leaves, where c_δ depends on δ. This is indeed the case for all $\delta \leq 5$. Kleitman and West [28] and Linial and Sturtevant [30] showed that every connected undirected graph G on n vertices with minimum degree at least 3 has a spanning tree with at least $n/4+2$ leaves. Griggs and Wu [24] proved that the maximum number of leaves in a spanning tree is at least $n/2+2$ when $\delta = 5$ and at least $2n/5+8/5$ when $\delta = 4$. All these results are tight. The situation is less clear for $\delta \geq 6$; the first author observed that Linial's conjecture is false for all large values of δ. Indeed, the results in [1] imply that there are undirected graphs with n vertices and minimum degree δ in which no tree has more than $(1-(1+o(1))\frac{\ln(\delta+1)}{\delta+1})n$ leaves, where the $o(1)$-term tends to zero and δ tends to infinity, and this is essentially tight. See also [4, pp. 4–5] and [11] for more information.

In this paper we prove an analogue of the Kleitman–West result for directed graphs: Every strongly connected digraph D of order n with minimum in-degree at least 3 has an out-branching with at least $(n/4)^{1/3}-1$ leaves. Unlike in the case of symmetric digraphs, in the case of all strongly connected digraphs, there is no linear lower bound: We show that there are strongly connected digraphs with minimum in-degree 3 in which every out-branching has at most $O(\sqrt{n})$ leaves.

Unlike its undirected counterpart which has attracted a lot of attention in all algorithmic paradigms like approximation algorithms [22, 31, 33], parameterized algorithms [9, 17, 18], exact exponential time algorithms [21], and also combinatorial studies [14, 24, 28, 30], the Directed Maximum Leaf Out-Branching problem has been neglected until the appearance of our conference papers [2] and [3].

Our second combinatorial result relates the number of leaves in a DMLOB of a directed graph D with the pathwidth of its underlying graph UG(D). (We postpone the definition of pathwidth till the next section.) If an undirected graph G contains a star $K_{1,k}$ as a minor, then it is possible to construct a spanning tree with at least k leaves from this minor. Otherwise, there is no $K_{1,k}$ minor in G, and it is possible to prove that the pathwidth of G is $O(k)$. (See, e.g., [8].) Actually, a much more general result due to Bienstock et al. [7] is that any undirected graph of pathwidth at least k contains all trees on k vertices as a minor. We prove a result that can be viewed as a generalization of known bounds on the number of leaves in a spanning tree of an undirected graph in terms of its pathwidth to strongly connected digraphs. We show that either a strongly connected digraph D has a DMLOB with at least k leaves or the pathwidth of UG(D) is $O(k \log k)$. For an acyclic digraph with a DMLOB having k leaves, we prove that the pathwidth is at most $4k$. This almost matches the bound for undirected graphs. These combinatorial results are useful in the design of parameterized algorithms.

In parameterized algorithms, for decision problems with input size n and a parameter k, the goal is to design an algorithm with runtime $f(k)n^{O(1)}$, where f is a function of k alone. (For DMLOB such a parameter is the number of leaves in the out-tree.) Problems having such an algorithm are said to be fixed parameter tractable (FPT). The book by Downey and Fellows [15] provides an introduction to the topic of parameterized complexity. For recent developments see the books by Flum and

Grohe [20] and by Niedermeier [32].

The parameterized version of DMLOB is defined as follows: Given a digraph D and a positive integral parameter k, does D contain an out-branching with at least k leaves? We denote the parameterized versions of DMLOB by k-DMLOB. If in the above definition we do not insist on an out-branching and ask whether there exists an out-tree with at least k leaves, we get the parameterized Directed Maximum Leaf Out-Tree problem (denoted k-DMLOT).

Our combinatorial bounds, combined with dynamic programming on graphs of bounded pathwidth, imply the first parameterized algorithms for k-DMLOB on strongly connected digraphs and acyclic digraphs. We remark that the algorithmic results presented here also hold for all digraphs if we consider k-DMLOT rather than k-DMLOB. This answers an open question of Fellows [12, 19, 25]. However, we restrict ourselves mainly to k-DMLOB for clarity and the harder challenges it poses, and we briefly consider k-DMLOT only in the last section.

This paper is organized as follows. In section 2 we provide additional terminology and notation as well as some well-known results. We introduce locally optimal out-branchings in section 3. Bounds on the number of leaves in maximum leaf out-branchings of strongly connected and acyclic digraphs are obtained in section 4. In section 5 we prove upper bounds on the pathwidth of the underlying graph of strongly connected and acyclic digraphs that do not contain out-branchings with at least k leaves. In section 6 we show that k-DMLOT is FPT. We give a brief overview of further research triggered by our papers [2] and [3] in section 7.

2. Preliminaries. Let D be a digraph. By $V(D)$ and $A(D)$ we represent the vertex set and arc set of D, respectively. An *oriented graph* is a digraph with no directed 2-cycle. Given a subset $V' \subseteq V(D)$ of a digraph D, let $D[V']$ denote the digraph induced by V'. The *underlying graph* $\mathrm{UG}(D)$ of D is obtained from D by omitting all orientations of arcs and by deleting one edge from each resulting pair of parallel edges. The *connectivity components* of D are the subdigraphs of D induced by the vertices of components of $\mathrm{UG}(D)$. A digraph D is *strongly connected* if, for every pair x, y of vertices, there are directed paths from x to y and from y to x. A maximal strongly connected subdigraph of D is called a *strong component*. A vertex u of D is an *in-neighbor* (*out-neighbor*) of a vertex v if $uv \in A(D)$ ($vu \in A(D)$, respectively). The *in-degree* $d^-(v)$ (*out-degree* $d^+(v)$) of a vertex v is the number of its in-neighbors (out-neighbors).

We denote by $\ell(D)$ the maximum number of leaves in an out-tree of a digraph D and by $\ell_s(D)$ we denote the maximum possible number of leaves in an out-branching of a digraph D. When D has no out-branching, we write $\ell_s(D) = 0$. The following simple result gives necessary and sufficient conditions for a digraph to have an out-branching. This assertion allows us to check whether $\ell_s(D) > 0$ in time $O(|V(D)| + |A(D)|)$.

Proposition 1 (see [6]). *A digraph D has an out-branching if and only if D has a unique strong component with no incoming arcs.*

Let $P = u_1u_2 \dots u_q$ be a directed path in a digraph D. An arc u_iu_j of D is a *forward* (*backward*) *arc for* P if $i \leq j-2$ ($j < i$, respectively). Every backward arc of the type $v_{i+1}v_i$ is called *double*.

For a natural number n, $[n]$ denotes the set $\{1, 2, \dots, n\}$.

A *tree decomposition* of an (undirected) graph G is a pair (X, U) where U is a tree whose vertices we will call *nodes* and $X = (\{X_i \mid i \in V(U)\})$ is a collection of subsets of $V(G)$ such that

1. $\bigcup_{i \in V(U)} X_i = V(G)$,

2. for each edge $\{v, w\} \in E(G)$, there is an $i \in V(U)$ such that $v, w \in X_i$, and
3. for each $v \in V(G)$ the set of nodes $\{i \mid v \in X_i\}$ forms a subtree of U.

The *width* of a tree decomposition $(\{X_i \mid i \in V(U)\}, U)$ equals $\max_{i \in V(U)}\{|X_i| - 1\}$. The *treewidth* of a graph G is the minimum width over all tree decompositions of G.

If in the definitions of a tree decomposition and treewidth we restrict U to be a path, then we have the definitions of path decomposition and pathwidth. We use the notation $tw(G)$ and $pw(G)$ to denote the treewidth and the pathwidth of a graph G.

We also need an equivalent definition of pathwidth in terms of vertex separators with respect to a linear ordering of the vertices. Let G be a graph and let $\sigma = (v_1, v_2, \ldots, v_n)$ be an ordering of $V(G)$. For $j \in [n]$ put $V_j = \{v_i : \ i \in [j]\}$ and denote by ∂V_j all vertices of V_j that have neighbors in $V \setminus V_j$. Setting $vs(G, \sigma) = \max_{i \in [n]} |\partial V_i|$, we define the *vertex separation* of G as

$$vs(G) = \min\{vs(G, \sigma) \colon \sigma \text{ is an ordering of } V(G)\}.$$

The following assertion is well known. It follows directly from the results of Kirousis and Papadimitriou [27] on interval width of a graph; see also [26].

PROPOSITION 2 (see [26, 27]). *For any graph G, $vs(G) = pw(G)$.*

3. Locally optimal out-branchings. Our bounds are based on finding locally optimal out-branchings. Given a digraph D and an out-branching T, we call a vertex *leaf*, *link*, or *branch* if its out-degree in T is 0, 1, or ≥ 2 respectively. Let $S^+_{\geq 2}(T)$ be the set of branch vertices, $S^+_1(T)$ the set of link vertices, and $L(T)$ the set of leaves in the tree T. Let $\mathscr{P}_2(T)$ be the set of maximal paths consisting of link vertices. By $p(v)$ we denote the *parent* of a vertex v in T; $p(v)$ is the unique in-neighbor of v. We call a pair of vertices u and v *siblings* if they do not belong to the same path from the root r in T. We start with the following well-known and easy to observe facts.

Fact 1. $|S^+_{\geq 2}(T)| \leq |L(T)| - 1$.

Fact 2. $|\mathscr{P}_2(T)| \leq 2|L(T)| - 1$.

Now we define the notion of local exchange which is intensively used in our proofs.

DEFINITION 3. ℓ-ARC EXCHANGE (ℓ-AE) OPTIMAL OUT-BRANCHING: *An out-branching T of a directed graph D with k leaves is ℓ-AE optimal if, for all arc subsets $F \subseteq A(T)$ and $X \subseteq A(D) - A(T)$ of size ℓ, $(A(T) \setminus F) \cup X$ is either not an out-branching, or an out-branching with at most k leaves. In other words, T is ℓ-AE optimal if it cannot be turned into an out-branching with more leaves by exchanging ℓ arcs.*

Let us remark that, for every fixed ℓ, an ℓ-AE optimal out-branching can be obtained in polynomial time. In our proofs we use only 1-AE optimal out-branchings. We need the following simple properties of 1-AE optimal out-branchings.

LEMMA 1. *Let T be a 1-AE optimal out-branching rooted at r in a digraph D. Then the following hold:*

(a) *For every pair of siblings $u, v \in V(T) \setminus L$ with $d^+_T(p(v)) = 1$, there is no arc $e = (u, v) \in A(D) \setminus A(T)$.*
(b) *For every pair of vertices $u, v \notin L$, $d^+_T(p(v)) = 1$, which are on the same path from the root with $dist(r, u) < dist(r, v)$ there is no arc $e = (u, v) \in A(D) \setminus A(T)$ (here $dist(r, u)$ is the distance to u in T from the root r).*
(c) *There is no arc (v, r), $v \notin L$, such that the directed cycle formed by the (r, v)-path and the arc (v, r) contains a vertex x such that $d^+_T(p(x)) = 1$.*

Proof. The proof easily follows from the fact that the existence of any of these arcs contradicts the local optimality of T with respect to 1-AE. ∎

4. Combinatorial bounds. We start with a lemma that allows us to obtain lower bounds on $\ell_s(D)$.

LEMMA 2. *Let D be an oriented graph of order n in which every vertex is of in-degree 2, and let D have an out-branching. If D has no out-tree with k leaves, then $n \leq 4k^3$.*

Proof. Let us assume that D has no out-tree with k leaves. Consider an out-branching T of D with $p < k$ leaves which is 1-AE optimal. Let r be the root of T.

We will bound the number n of vertices in T as follows. Every vertex of T is either a leaf, or a branch vertex, or a link vertex. By Facts 1 and 2 we already have bounds on the number of leaf and branch vertices as well as the number of maximal paths consisting of link vertices. So to get an upper bound on n in terms of k, it suffices to bound the length of each maximal path consisting of link vertices. Let us consider such a path P and let x, y be the first and last vertices of P, respectively.

The vertices of $V(T) \setminus V(P)$ can be partitioned into four classes as follows:

(a) *ancestor vertices*: the vertices which appear before x on the (r, x)-path of T;
(b) *descendant vertices*: the vertices appearing after the vertices of P on paths of T starting at r and passing through y;
(c) *sink vertices*: the vertices which are leaves but not descendant vertices;
(d) *special vertices*: none-of-the-above vertices.

Let $P' = P - x$, let z be the out-neighbor of y on T, and let T_z be the subtree of T rooted at z. By Lemma 1, there are no arcs from special or ancestor vertices to the path P'. Let uv be an arc of $A(D) \setminus A(P')$ such that $v \in V(P')$. There are two possibilities for u: (i) $u \notin V(P')$ or (ii) $u \in V(P')$ and uv is backward for P' (there are no forward arcs for P' since T is 1-AE optimal). Note that every vertex of type (i) is either a descendant vertex or a sink. Since every vertex of D is of in-degree 2, the backward arcs for P' form a vertex-disjoint collection of out-trees with roots at vertices that are not terminal vertices of backward arcs for P'. These roots are terminal vertices of arcs in which first vertices are descendant vertices or sinks.

We denote by $\{u_1, u_2, \ldots, u_s\}$ and $\{v_1, v_2, \ldots, v_t\}$ the sets of vertices on P' which have in-neighbors that are descendant vertices and sinks, respectively. Let the out-tree formed by backward arcs for P' rooted at $w \in \{u_1, \ldots, u_s, v_1, \ldots, v_t\}$ be denoted by $T(w)$, and let $l(w)$ denote the number of leaves in $T(w)$. Observe that the following is an out-tree rooted at z:

$$T_z \cup \{(in(u_1), u_1), \ldots, (in(u_s), u_s)\} \cup \bigcup_{i=1}^{s} T(u_i),$$

where $\{in(u_1), \ldots, in(u_s)\}$ are the in-neighbors of $\{u_1, \ldots, u_s\}$ on T_z. This out-tree has at least $\sum_{i=1}^{s} l(u_i)$ leaves, and, thus, $\sum_{i=1}^{s} l(u_i) \leq k-1$. Let us denote the subtree of T rooted at x by T_x and let $\{in(v_1), \ldots, in(v_t)\}$ be the in-neighbors of $\{v_1, \ldots, v_t\}$ on $T - V(T_x)$. Then we have the following out-tree:

$$(T - V(T_x)) \cup \{(in(v_1), v_1), \ldots, (in(v_t), v_t)\} \cup \bigcup_{i=1}^{t} T(v_i)$$

with at least $\sum_{i=1}^{t} l(v_i)$ leaves. Thus, $\sum_{i=1}^{t} l(v_i) \leq k - 1$.

Consider a path $R = p_0 p_1 \ldots p_r$ formed by backward arcs. Observe that the arcs $\{p_i p_{i+1} : 0 \leq i \leq r-1\} \cup \{p_j p_j^+ : 1 \leq j \leq r\}$ form an out-tree with r leaves, where p_j^+ is the out-neighbor of p_j on P. Thus, there is no path of backward arcs of

length more than $k-1$. Every out-tree $T(w)$, $w \in \{u_1, \ldots, u_s\}$, has $l(w)$ leaves, and, thus, its arcs can be decomposed into $l(w)$ paths, each of length at most $k-1$. Now we can bound the number of arcs in all the trees $T(w)$, $w \in \{u_1, \ldots, u_s\}$, as follows: $\sum_{i=1}^{s} l(u_i)(k-1) \leq (k-1)^2$. We can similarly bound the number of arcs in all the trees $T(w)$, $w \in \{v_1, \ldots, v_s\}$, by $(k-1)^2$. Recall that the vertices of P' can be either terminal vertices of backward arcs for P' or vertices in $\{u_1, \ldots, u_s, v_1, \ldots, v_t\}$. Observe that $s + t \leq 2(k-1)$ since $\sum_{i=1}^{s} l(u_i) \leq k-1$ and $\sum_{i=1}^{t} l(v_i) \leq k-1$.

Thus, the number of vertices in P is bounded from above by $1+2(k-1)+2(k-1)^2$. Therefore,

$$\begin{aligned} n &= |L(T)| + |S^+_{\geq 2}(T)| + |S^+_1(T)| \\ &= |L(T)| + |S^+_{\geq 2}(T)| + \sum_{P \in \mathscr{P}_2(T)} |V(P)| \\ &\leq (k-1) + (k-2) + (2k-3)(2k^2 - 2k + 1) \\ &< 4k^3. \end{aligned}$$

Thus, we conclude that $n \leq 4k^3$. ∎

THEOREM 4. *Let D be a strongly connected digraph with n vertices.*

(a) *If D is an oriented graph with minimum in-degree at least 2, then $\ell_s(D) \geq (n/4)^{1/3} - 1$.*

(b) *If D is a digraph with minimum in-degree at least 3, then $\ell_s(D) \geq (n/4)^{1/3} - 1$.*

Proof. Since D is strongly connected, we have $\ell(D) = \ell_s(D) > 0$. Let T be a 1-AE optimal out-branching of D with maximum number of leaves. (a) Delete some arcs from $A(D) \setminus A(T)$, if needed, such that the in-degree of each vertex of D becomes 2. Now the inequality $\ell_s(D) \geq (n/4)^{1/3} - 1$ follows from Lemma 2 and the fact that $\ell(D) = \ell_s(D)$.

(b) Let P be the path formed in the proof of Lemma 2. (Note that $A(P) \subseteq A(T)$.) Delete every double arc of P, in case there are any, and delete some more arcs from $A(D) \setminus A(T)$, if needed, to ensure that the in-degree of each vertex of D becomes 2. It is not difficult to see that the proof of Lemma 2 remains valid for the new digraph D. Now the inequality $\ell_s(D) \geq (n/4)^{1/3} - 1$ follows from Lemma 2 and the fact that $\ell(D) = \ell_s(D)$. ∎

Remark 5. It is easy to see that Theorem 4 also holds for acyclic digraphs D with $\ell_s(D) > 0$.

While we do not know whether the bounds of Theorem 4 are tight, we can show that no linear bounds are possible. The following result is formulated for part (b) of Theorem 4, but a similar result holds for part (a) as well.

THEOREM 6. *For each $t \geq 6$ there is a strongly connected digraph H_t of order $n = t^2 + 1$ with minimum in-degree 3 such that $0 < \ell_s(H_t) = O(t)$.*

Proof. Let $V(H_t) = \{r\} \cup \{u^i_1, u^i_2, \ldots, u^i_t \mid i \in [t]\}$ and

$$\begin{aligned} A(H_t) = &\left\{u^i_j u^i_{j+1}, u^i_{j+1} u^i_j \mid \ i \in [t],\ j \in \{0, 1, \ldots, t-4\}\right\} \\ &\bigcup \left\{u^i_j u^i_{j-2} \mid i \in [t],\ j \in \{3, 4, \ldots, t-2\}\right\} \\ &\bigcup \left\{u^i_j u^i_q \mid \ i \in [t],\ t-3 \leq j \neq q \leq t\right\}, \end{aligned}$$

where $u^i_0 = r$ for every $i \in [t]$. It is easy to check that $0 < \ell_s(H_t) = O(t)$. ∎

5. Pathwidth of underlying graphs and parameterized algorithms. By Proposition 1, an acyclic digraph D has an out-branching if and only if D possesses a single vertex of in-degree zero.

THEOREM 7. *Let D be an acyclic digraph with a single vertex of in-degree zero. Then either $\ell_s(D) \geq k$ or the underlying undirected graph of D is of pathwidth at most $4k$ and we can obtain this path decomposition in polynomial time.*

Proof. Assume that $\ell_s(D) \leq k-1$. Consider a 1-AE optimal out-branching T of D. Notice that $|L(T)| \leq k-1$. Now remove all the leaves and branch vertices from the tree T. The remaining vertices form maximal directed paths consisting of link vertices. Delete the first vertices of all paths. As a result we obtain a collection $\mathcal{Q}$ of directed paths. Let $H = \cup_{P \in \mathcal{Q}} P$. We will show that every arc uv with $u, v \in V(H)$ is in H. Let $P' \in \mathcal{Q}$. As in the proof of Lemma 2, we see that there are no forward arcs for P'. Since D is acyclic, there are no backward arcs for P'.

Suppose uv is an arc of D such that $u \in R'$ and $v \in P'$, where R' and P' are distinct paths from $\mathcal{Q}$. As in the proof of Lemma 2, we see that u is either a sink or a descendent vertex for P' in T. Since R' contains no sinks of T, u is a descendent vertex, which is impossible as D is acyclic. Thus, we have proved that $pw(\mathrm{UG}(H)) = 1$.

Consider a path decomposition of H of width 1. We can obtain a path decomposition of $\mathrm{UG}(D)$ by adding all the vertices of $L(T) \cup S^+_{\geq 2}(T) \cup F(T)$, where $F(T)$ is the set of first vertices of maximal directed paths consisting of link vertices of T, to each of the bags of a path decomposition of H of width 1. Observe that the pathwidth of this decomposition is bounded from above by

$$|L(T)| + |S^+_{\geq 2}(T)| + |F(T)| + 1 \leq (k-1) + (k-2) + (2k-3) + 1 \leq 4k - 5.$$

The bounds on the various sets in the inequality above follows from Facts 1 and 2. This proves the theorem. ∎

COROLLARY 1. *For acyclic digraphs, the problem k-DMLOB can be solved in time $2^{O(k \log k)} \cdot n^{O(1)}$.*

Proof. The proof of Theorem 7 can be easily turned into a polynomial time algorithm to either build an out-branching of D with at least k leaves or show that $pw(\mathrm{UG}(D)) \leq 4k$ and provide the corresponding path decomposition. A standard dynamic programming over the path (tree) decomposition (see, e.g., [5]) gives us an algorithm of running time $2^{O(k \log k)} \cdot n^{O(1)}$. ∎

The following simple lemma is well known; see, e.g., [13].

LEMMA 3. *Let $T = (V, E)$ be an undirected tree and let $w : V \to \mathbb{R}^+ \cup \{0\}$ be a weight function on its vertices. There exists a vertex $v \in T$ such that the weight of every subtree T' of $T - v$ is at most $w(T)/2$, where $w(T) = \sum_{v \in V} w(v)$.*

Let D be a strongly connected digraph and let T be an out-branching of D with λ leaves. Consider the following decomposition of T (called a *β-decomposition*) which will be useful in the proof of Theorem 8.

Assign weight 1 to all leaves of T and weight 0 to all nonleaves of T. By Lemma 3, T has a vertex v such that each component of $T - v$ has at most $\lambda/2 + 1$ leaves (if v is not the root and its in-neighbor v^- in T is a link vertex, then v^- becomes a new leaf). Let $T_1, T_2, \ldots, T_s$ be the components of $T - v$ and let $l_1, l_2, \ldots, l_s$ be the numbers of leaves in the components. Notice that $\lambda \leq \sum_{i=1}^{s} l_i \leq \lambda + 1$ (we may get a new leaf). We may assume that $l_s \leq l_{s-1} \leq \cdots \leq l_1 \leq \lambda/2 + 1$. Let j be the smallest index such that $\sum_{i=1}^{j} l_i \geq \frac{\lambda}{2} + 1$. Consider two cases: (a) $l_j \leq (\lambda + 2)/4$ and (b) $l_j > (\lambda + 2)/4$.

In case (a), we have

$$\frac{\lambda+2}{2} \leq \sum_{i=1}^{j} l_i \leq \frac{3(\lambda+2)}{4} \quad \text{and} \quad \frac{\lambda-6}{4} \leq \sum_{i=j+1}^{s} l_i \leq \frac{\lambda}{2}.$$

In case (b), we have $j = 2$ and

$$\frac{\lambda+2}{4} \leq l_1 \leq \frac{\lambda+2}{2} \quad \text{and} \quad \frac{\lambda-2}{2} \leq \sum_{i=2}^{s} l_i \leq \frac{3\lambda+2}{4}.$$

Let $p = j$ in case (a) and $p = 1$ in case (b). Add to D and T a *copy* v' of v (with the same in- and out-neighbors). Then the number of leaves in each of the out-trees

$$T' = T[\{v\} \cup (\cup_{i=1}^{p} V(T_i))] \quad \text{and} \quad T'' = T[\{v'\} \cup (\cup_{i=p+1}^{s} V(T_i))]$$

is between $\lambda(1+o(1))/4$ and $3\lambda(1+o(1))/4$. Observe that the vertices of T' have at most $\lambda+1$ out-neighbors in T'' and the vertices of T'' have at most $\lambda+1$ out-neighbors in T' (we add 1 to λ due to the fact that v "belongs" to both T' and T'').

Similarly to deriving T' and T'' from T, we can obtain two out-trees from T' and two out-trees from T'' in which the numbers of leaves are approximately between a quarter and three quarters of the number of leaves in T' and T'', respectively. Observe that after $O(\log \lambda)$ "dividing" steps, we will end up with $O(\lambda)$ out-trees with just one leaf, i.e., directed paths. These paths contain $O(\lambda)$ copies of vertices of D (such as v' above). After deleting the copies, we obtain a collection of $O(\lambda)$ disjoint directed paths covering $V(D)$.

THEOREM 8. *Let D be a strongly connected digraph. Then either $\ell_s(D) \geq k$ or the underlying undirected graph of D is of pathwidth $O(k \log k)$.*

Proof. We may assume that $\ell_s(D) < k$. Let T be a 1-AE optimal out-branching, and let λ be the number of leaves in T. Consider a β-decomposition of T. The decomposition process can be viewed as a tree $\mathcal{T}$ rooted in a node (associated with) T. The children of T in $\mathcal{T}$ are nodes (associated with) T' and T''; the leaves of $\mathcal{T}$ are the directed paths of the decomposition. The *first layer* of $\mathcal{T}$ is the node T, the *second layer* consists of T' and T'', the *third layer* consists of the children of T' and T'', etc. In what follows, we do not distinguish between a node Q of $\mathcal{T}$ and the tree associated with the node. Assume that $\mathcal{T}$ has t layers. Notice that the last layer consists of (some) leaves of $\mathcal{T}$ and that $t = O(\log k)$, which was proved above (note that $\lambda \leq k-1$).

Let Q be a node of $\mathcal{T}$ at layer j. We will prove that

$$pw(\mathrm{UG}(D[V(Q)])) < 2(t-j+2.5)k. \tag{1}$$

Since $t = O(\log k)$, (1) for $j = 1$ implies that the underlying undirected graph of D is of pathwidth $O(k \log k)$.

We first prove (1) for $j = t$ when Q is a path from the decomposition. Let $W = (L(T) \cup S^{+}_{\geq 2}(T) \cup F(T)) \cap V(Q)$, where $F(T)$ is the set of first vertices of maximal paths of T consisting of link vertices. As in the proof of Theorem 7, it follows from Facts 1 and 2 that $|W| < 4k$. Obtain a digraph R by deleting from $D[V(Q)]$ all arcs in which at least one end-vertex is in W and which are not arcs of Q. As in the proof of Theorem 7, it follows from Lemma 1 and 1-AE optimality of T that there are no forward arcs for Q in R. Let $Q = v_1 v_2 \dots v_q$. For every $j \in [q]$, let $V_j = \{v_i : \ i \in [j]\}$.

If for some j the set V_j contained k vertices, say $\{v'_1, v'_2, \ldots, v'_k\}$, having in-neighbors in the set $\{v_{j+1}, v_{j+2}, \ldots, v_q\}$, then D would contain an out-tree with k leaves formed by the path $v_{j+1}v_{j+2}\ldots v_q$ together with a backward arc terminating at v'_i from a vertex on the path for each $1 \leq i \leq k$, a contradiction. Thus $vs(\mathrm{UG}(D_2[P])) \leq k$. By Proposition 2, the pathwidth of $\mathrm{UG}(R)$ is at most k. Let $(X_1, X_2, \ldots, X_s)$ be a path decomposition of $\mathrm{UG}(R)$ of width at most k. Then $(X_1 \cup W, X_2 \cup W, \ldots, X_s \cup W)$ is a path decomposition of $\mathrm{UG}(D[V(Q)])$ of width less than $k + 4k$. Thus,

$$pw(\mathrm{UG}(D[V(Q)])) < 5k. \tag{2}$$

Now assume that we have proved (1) for $j = i$ and show it for $j = i-1$. Let Q be a node of layer $i-1$. If Q is a leaf of $\mathcal{T}$, we are done by (2). Thus, we may assume that Q has children Q' and Q'' which are nodes of layer i. In the β-decomposition of T given before this theorem, we saw that the vertices of T' have at most $\lambda + 1$ out-neighbors in T'' and the vertices of T'' have at most $\lambda + 1$ out-neighbors in T'. Similarly, we can see that (in the β-decomposition of this proof) the vertices of Q' have at most k out-neighbors in Q'' and the vertices of Q'' have at most k out-neighbors in Q' (since $\lambda \leq k-1$). Let Y denote the set of the above-mentioned out-neighbors on Q' and Q''; $|Y| \leq 2k$. Delete from $D[V(Q') \cup V(Q'')]$ all arcs in which at least one end-vertex is in Y and which do not belong to $Q' \cup Q''$.

Let G denote the obtained digraph. Observe that G is disconnected and $G[V(Q')]$ and $G[V(Q'')]$ are components of G. Thus, $pw(\mathrm{UG}(G)) \leq b$, where

$$b = \max\{pw(\mathrm{UG}(G[V(Q')])), pw(\mathrm{UG}(G[V(Q'')]))\} < 2(t - i + 2.5)k. \tag{3}$$

Let $(Z_1, Z_2, \ldots, Z_r)$ be a path decomposition of G of width at most b. Then $(Z_1 \cup Y, Z_2 \cup Y, \ldots, Z_r \cup Y)$ is a path decomposition of $\mathrm{UG}(D[V(Q') \cup V(Q'')])$ of width at most $b + 2k < 2(t - (i-1) + 2.5)k$. This completes the proof. □

Similarly to the proof of Corollary 1, we obtain the following corollary.

Corollary 2. *For a strongly connected digraph D, the problem k-DMLOB can be solved in time $2^{O(k \log^2 k)} \cdot n^{O(1)}$.*

6. k-DMLOT is FPT. Observe that while our results are for strongly connected digraphs, they can be extended to a larger class of digraphs. Notice that $\ell(D) \geq \ell_s(D)$ for each digraph D. Let $\mathcal{L}$ be the family of digraphs D for which either $\ell_s(D) = 0$ or $\ell_s(D) = \ell(D)$. The following assertion shows that $\mathcal{L}$ includes a large number of digraphs including all strongly connected digraphs and acyclic digraphs (and, also, the well-studied classes of semicomplete multipartite digraphs, and quasi-transitive digraphs; see [6] for the definitions).

Proposition 3 (see [2]). *Suppose that a digraph D satisfies the following property: for every pair R and Q of distinct strong components of D, if there is an arc from R to Q, then each vertex of Q has an in-neighbor in R. Then $D \in \mathcal{L}$.*

Let $\mathcal{B}$ be the family of digraphs that contain out-branchings. The results of this paper proved for strongly connected digraphs can be extended to the class $\mathcal{L} \cap \mathcal{B}$ of digraphs since in the proofs we use only the following property of strongly connected digraphs D: $\ell_s(D) = \ell(D) > 0$.

For a digraph D and a vertex v, let D_v denote the subdigraph of D induced by all vertices reachable from v. Using the $2^{O(k \log^2 k)} \cdot n^{O(1)}$ algorithm for k-DMLOB on digraphs in $\mathcal{L} \cap \mathcal{B}$ and the facts that (i) $D_v \in \mathcal{L} \cap \mathcal{B}$ for each digraph D and vertex v and (ii) $\ell(D) = \max\{\ell_s(D_v) | v \in V(D)\}$ (for details, see [2]), we can obtain an $2^{O(k \log^2 k)} \cdot n^{O(1)}$ algorithm for k-DMLOT on *all* digraphs. For acyclic digraphs, the running time can be reduced to $2^{O(k \log k)} \cdot n^{O(1)}$.

7. Consequent research. Research initiated by [2] and [3] was continued by Bonsma and Dorn who proved in [10] that every strongly connected digraph of order n with minimum in-degree at least 3 has an out-branching with at least $\sqrt{n}/4$ leaves. Thus, the maximum guaranteed number $\lambda(n)$ of leaves in a strongly connected digraph of order n with minimum in-degree at least 3 is $\Theta(\sqrt{n})$. It would be interesting to obtain the maximum constant c such that $\lambda(n) \geq c\sqrt{n}$.

Using several ideas of this paper, some new ideas and treewidth rather than pathwidth, Bonsma and Dorn [10] designed algorithms of complexity $2^{O(k \log k)} n^{O(1)}$ for both k-DMLOT and k-DMLOB. Using another approach, Kneis, Langer, and Rossmanith [29] obtained a $4^k n^{O(1)}$ time algorithm for k-DMLOB. It is not difficult to see that this algorithm implies an $4^k n^{O(1)}$ time algorithm for k-DMLOT.

We conclude by pointing out that in a recent paper [16], Drescher and Vetta describe an $O(\sqrt{\text{OPT}})$-approximation algorithm for DMLOB, where OPT is the maximum number of leaves in an out-branching of the input digraph.

Acknowledgment. We would like to thank the referees for a number of useful suggestions.

REFERENCES

[1] N. Alon, *Transversal numbers of uniform hypergraphs*, Graphs Combin., 6 (1990), pp. 1–4.
[2] N. Alon, F. V. Fomin, G. Gutin, M. Krivelevich, and S. Saurabh, *Parameterized algorithms for directed maximum leaf problems*, in Automata, Languages and Programming (ICALP 2007), Lecture Notes in Comput. Sci. 4596, Springer-Verlag, Berlin, 2007, pp. 352–362.
[3] N. Alon, F. V. Fomin, G. Gutin, M. Krivelevich, and S. Saurabh, *Better algorithms and bounds for directed maximum leaf problems*, in FSTTCS 2007: Foundations of Software Technology and Theoretical Computer Science, Lecture Notes in Comput. Sci. 4855, Springer-Verlag, Berlin, 2007, pp. 316–327.
[4] N. Alon and J. Spencer, *The Probabilistic Method*, 2nd ed., John Wiley & Sons, New York, 2000.
[5] S. Arnborg and A. Proskurowski, *Linear time algorithms for NP-hard problems restricted to partial k-trees*, Discrete Appl. Math., 23 (1989), pp. 11–24.
[6] J. Bang-Jensen and G. Gutin, *Digraphs. Theory, Algorithms and Applications*, Springer-Verlag, London, 2001.
[7] D. Bienstock, N. Robertson, P. D. Seymour, and R. Thomas, *Quickly excluding a forest*, J. Combin. Theory Ser. B, 52 (1991), pp. 274–283.
[8] H. L. Bodlaender, *On linear time minor tests and depth-first search*, J. Algorithms, 14 (1993), pp. 1–23.
[9] P. S. Bonsma, T. Brueggermann, and G. J. Woeginger, *A faster FPT algorithm for finding spanning trees with many leaves*, in Mathematical Foundations of Computer Science 2003, Lecture Notes in Comput. Sci. 2747, Springer-Verlag, Berlin, 2003, pp. 259–268.
[10] P. S. Bonsma and F. Dorn, *Tight bounds and faster algorithms for directed max-leaf*, in Proceedings of the 16th European Symposium on Algorithms, Lecture Notes in Comput. Sci. 5193, Springer-Verlag, Berlin, 2008, pp. 222-233.
[11] Y. Caro, D. B. West, and R. Yuster, *Connected domination and spanning trees with many leaves*, SIAM J. Discrete Math., 13 (2000), pp. 202–211.
[12] M. Cesati, *Compendium of Parameterized Problems*, http://bravo.ce.uniroma2.it/home/cesati/research/compendium.pdf, 2006.
[13] F. R. K. Chung, *Separator theorems and their applications*, in Paths, Flows, and VLSI-Layout (Bonn, 1988), Algorithms Combin. 9, Springer-Verlag, Berlin, 1990, pp. 17–34.
[14] G. Ding, Th. Johnson, and P. Seymour, *Spanning trees with many leaves*, J. Graph Theory, 37 (2001), pp. 189–197.
[15] R. G. Downey and M. R. Fellows, *Parameterized Complexity*, Springer-Verlag, New York, 1999.
[16] M. Drescher and A. Vetta, *An approximation algorithm for the maximum leaf spanning arborescence problem*, ACM Trans. Algorithms, to appear.

[17] V. Estivill-Castro, M. R. Fellows, M. A. Langston, and F. A. Rosamond, *FPT is P-time extremal structure* I, in Proceedings of ACiD, College Publications, London, 2005, pp. 1–41.
[18] M. R. Fellows, C. McCartin, F. A. Rosamond, and U. Stege, *Coordinated kernels and catalytic reductions: An improved FPT algorithm for max leaf spanning tree and other problems*, in Proceedings of the 20th Conference on Foundations of Software Technology and Theoretical Computer Science, Lecture Notes in Comput. Sci. 1974, Springer-Verlag, Berlin, 2000, pp. 240–251.
[19] M. Fellows, *private communications*, 2005–2006.
[20] J. Flum and M. Grohe, *Parameterized Complexity Theory*, Springer-Verlag, Berlin, 2006.
[21] F. V. Fomin, F. Grandoni, and D. Kratsch, *Solving connected dominating set faster than* 2^n, Algorithmica, 52 (2008), pp. 153–166.
[22] G. Galbiati, A. Morzenti, and F. Maffioli, *On the approximability of some maximum spanning tree problems*, Theoret. Comput. Sci., 181 (1997), pp. 107–118.
[23] M. R. Garey and D. S. Johnson, *Computers and Intractability*, W.H. Freeman, New York, 1979.
[24] J. R. Griggs and M. Wu, *Spanning trees in graphs of minimum degree four or five*, Discrete Math., 104 (1992), pp. 167–183.
[25] G. Gutin and A. Yeo, *Some parameterized problems on digraphs*, Comput. J., 51 (2008), pp. 363–371.
[26] N. G. Kinnersley, *The vertex separation number of a graph equals its path-width*, Inform. Process. Lett., 42 (1992), pp. 345–350.
[27] L. M. Kirousis and C. H. Papadimitriou, *Interval graphs and searching*, Discrete Math., 55 (1985), pp. 181–184.
[28] D. J. Kleitman and D. B. West, *Spanning trees with many leaves*, SIAM J. Discrete Math., 4 (1991), pp. 99–106.
[29] J. Kneis, A. Langer, and P. Rossmanith, *A new algorithm for finding trees with many leaves*, in Proceedings of the 19th International Symposium on Algorithms and Computation (ISAAC), Lecture Notes in Comput. Sci. 5369, Springer-Verlag, Berlin, 2008, pp. 270–281.
[30] N. Linial and D. Sturtevant, unpublished result, 1987.
[31] H.-I. Lu and R. Ravi, *Approximating maximum leaf spanning trees in almost linear time*, J. Algorithms, 29 (1998), pp. 132–141.
[32] R. Niedermeier, *Invitation to Fixed-Parameter Algorithms*, Oxford University Press, Oxford, UK, 2006.
[33] R. Solis-Oba, 2*-approximation algorithm for finding a spanning tree with maximum number of leaves*, in Algorithms—ESA '98, Lecture Notes in Comput. Sci. 1461, Springer-Verlag, Berlin, 1998, pp. 441–452.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 477–486

k-CHROMATIC NUMBER OF GRAPHS ON SURFACES*

ZDENĚK DVOŘÁK† AND RISTE ŠKREKOVSKI‡

Abstract. A well-known result (Heawood [*Quart. J. Pure Appl. Math.*, 24 (1890), pp. 332–338], Ringel [*Map Color Theorem*, Springer-Verlag, New York, 1974], Ringel and Youngs [*Proc. Nat. Acad. Sci.*, U.S.A., 60 (1968), pp. 438–445]) states that the maximum chromatic number of a graph embedded in a given surface S coincides with the size of the largest clique that can be embedded in S, and that this number can be expressed as a simple formula in the Euler genus of S. A *partition* of a graph G into k parts consists of k edge-disjoint subgraphs $G_1, \ldots, G_k$ such that $E(G) = E(G_1) \cup E(G_2) \cup \cdots \cup E(G_k)$. The *$k$-chromatic number* $\chi_k(G)$ is the maximum of $\sum_{i=1}^{k} \chi(G_i)$ over all partitions of G into k parts. We derive a Heawood-type formula for the k-chromatic number of graphs embedded in a fixed surface, improving the previously known upper bounds. In infinitely many cases, the new upper bound coincides with the lower bound obtained from embedding disjoint cliques in the surface. In the proof of this result, we derive a variant of Euler's formula for the union of several graphs that might be interesting independently.

Key words. graph decomposition, chromatic number, surface embedding, Euler's formula

AMS subject classifications. 05C15, 05C10

DOI. 10.1137/070688262

1. Introduction and definitions.

We consider simple undirected graphs with no loops and parallel edges. Let $e(G)$ and $n(G)$ denote the number of edges and the number of vertices of a graph G, respectively. When the graph G is clear from the context, we simply use e and n. A *proper coloring* of a graph G by k colors is assignment of colors $1, 2, \ldots, k$ to vertices of G such that no two adjacent vertices have the same color. The *chromatic number* $\chi(G)$ of graph G is the minimum k such that G has a proper coloring by k colors.

Let Σ_h denote the orientable surface obtained from the sphere by attaching h handles, and let Π_h be the nonorientable surface obtained from the sphere by inserting h crosscaps. The *Euler genus* $g(S)$ of a surface S is given by $g(\Sigma_h) = 2h$ and $g(\Pi_h) = h$. Let $g(G)$ denote the *Euler genus* of the graph G, i.e., the minimal Euler genus of a surface into which G is embeddable.

Colorings of graphs on surface have been studied extensively. The fundamental result in this area is the well-known Four Color Theorem that was proved by Appel and Haken [1] in 1977, and a shorter proof was later found by Robertson et al. [12]. Regarding the graphs on surfaces of genus $g \geq 1$, Heawood [6] showed that each graph embedded in such a surface has chromatic number at most

$$H(g) = \left\lfloor \frac{7 + \sqrt{24g+1}}{2} \right\rfloor.$$

Later, Ringel [11] and Ringel and Youngs [10] found the corresponding lower bounds by showing that the complete graph on $H(g)$ vertices can be embedded into any

*Received by the editors April 16, 2007; accepted for publication (in revised form) May 20, 2008; published electronically February 4, 2009. This research was supported in part by bilateral projects SLO-CZ/04-05-002 and MSMT-07-0405 between Slovenia and Czech Republic.
http://www.siam.org/journals/sidma/23-1/68826.html

†Faculty of Mathematics and Physics, Institute for Theoretical Computer Science (ITI), Charles University, Malostranské nám. 2/25, 118 00, Prague, Czech Republic (rakdver@kam.mff.cuni.cz).

‡Department of Mathematics, University of Ljubljana, Jadranska 19, 1111 Ljubljana, Slovenia (bluesky2high@yahoo.com).

surface of Euler genus g, with the exception of the Klein bottle, where the correct bound on the chromatic number is 6 (established by Franklin [4]).

We consider the properties (especially regarding the chromatic number) of partitions of a graph into several subgraphs. A *partition* of a graph G into k parts consists of k edge-disjoint subgraphs $G_1, \ldots, G_k$ such that $E(G) = E(G_1) \cup E(G_2) \cup \cdots \cup E(G_k)$. Note that we do not require that the subgraphs G_i be spanning, i.e., possibly $n(G_i) < n(G)$ for some i. We always assume that the graphs G_i do not contain isolated vertices. We call the subgraphs G_i parts of the partition.

The *k-chromatic number* $\chi_k(G)$ is the maximum of $\sum_{i=1}^{k} \chi(G_i)$ over all partitions $G_1, G_2, \ldots, G_k$ of G into k parts. The parameter χ_k has been studied for general graphs as well as for graphs of bounded genus. The fact that for a graph G with n vertices $\chi_2(G) \leq n + 1$ follows from the well-known theorem of Nordhaus and Gaddum [8]. Plesník [9] proved that $n + \binom{k}{2} \leq \chi_k(K_n) \leq n + 2^{\binom{k+1}{2}}$ and conjectured that $\chi_k(K_n) = n + \binom{k}{2}$. Watkinson [15] improved the upper bound to $\chi_k(K_n) \leq n + \frac{k!}{2}$ and Füredi et al. [5] to $\chi_k(K_n) \leq n + 7^k$.

Regarding the graphs with bounded genus, let us define $\chi_k(S)$ to be the maximum of $\chi_k(G)$ over all graphs G that can be embedded in the surface S. Stiebitz and Škrekovski [14] have determined the exact values of χ_2 for all surfaces. Füredi et al. [5] have shown that

$$\chi_k(S) \leq \left\lfloor \frac{7k + \sqrt{24kg + 49k^2 - 48k}}{2} \right\rfloor,$$

where g denotes the Euler genus of S. They also found a lower bound of order

$$\frac{7k + \sqrt{24kg + k^2}}{2}.$$

In this paper, we decrease the upper bound; this way, we obtain exact values for many surfaces and values of k.

THEOREM 1. *Let G be a simple graph G of Euler genus g. If $k \leq g$, then*

$$\chi_k(G) \leq \left\lfloor \frac{7k + \sqrt{24kg + k^2}}{2} \right\rfloor.$$

An embedding of a graph in a surface is called *cellular* if the interior of each face is homeomorphic to an open disk. In particular, the boundary walk of each face in a cellular embedding is connected. For a face f of such an embedding, let $\ell(f)$ be the length of its boundary walk. If G is a simple connected graph with at least three vertices, then $\ell(f) \geq 3$ for each face f. A *block* of a graph G is a maximum 2-connected induced subgraph of G. Let us recall some fundamental facts about graph embeddings and surfaces that can be found, e.g., in [7].

THEOREM 2. *Let G be a connected graph of Euler genus g. Then, any embedding of G in a surface with Euler genus g is cellular.*

THEOREM 3 (Battle et al. [2], Stahl and Beineke [13]). *If $G_1, G_2, \ldots, G_n$ are the blocks of a graph G, then*

$$g(G) = \sum_{i=1}^{n} g(G_i).$$

THEOREM 4 (Franklin [4], Ringel [11], Ringel and Youngs [10]). *The Euler genus of the complete graph K_n is $g = \lceil \frac{1}{6}(n-3)(n-4) \rceil$. K_n can be embedded into any*

surface with Euler genus g, with the exception of K_7, that cannot be embedded in Π_2, i.e., the Klein bottle.

THEOREM 5 (Euler's formula). *If f is the number of faces of a cellular embedding of a graph G into a surface of Euler genus g, then $e(G) = n(G) + f + g - 2$.*

In the following section, we derive a version of Euler's formula that provides more information about a graph split into several parts (Theorem 8).

A graph G is *critical* if for every proper subgraph $H \subset G$, $\chi(H) < \chi(G)$. If G is a critical graph and $\chi(G) = k$, we say that G is *k-critical*. Obviously, if G is k-critical, then $\delta(G) \geq k - 1$. For noncomplete graphs, the following stronger result known as Dirac's inequality was shown in [3].

THEOREM 6 (Dirac). *If G is a k-critical graph with $k \geq 4$ and G is not a clique, then $2e(G) \geq (k-1)n(G) + k - 3$.*

2. Generalized Euler's formula. Let F be the set of the faces of a cellular embedding of a simple connected graph G with at least three vertices. Then, $\Delta = \sum_{f \in F}(\ell(f) - 3) \geq 0$ is the number of edges that must be added to G to make it a triangulation (possibly introducing parallel edges and loops during the construction). One of the well-known consequences of Euler's formula is the following proposition.

PROPOSITION 7. *If G is a simple connected graph with $n \geq 3$ vertices and e edges embedded cellularly to a surface of Euler genus g, then $e + \Delta = 3n + 3g - 6$. In particular, $e \leq 3n + 3g(G) - 6$.*

We include the proof for the sake of completeness.

Proof. Let F be the set of faces of G. Since each edge of G appears exactly twice in the facial walks, we have $2e = \sum_{f \in F} \ell(f)$, and consequently $2e - \Delta = 3|F|$. Using Theorem 5, we infer that $3e = 3n + 3|F| + 3g - 6 = 3n + 2e - \Delta + 3g - 6$, from which the desired formula immediately follows. Also, by Theorem 2, the embedding of G into a surface of Euler genus $g(G)$ is cellular, and since $\Delta \geq 0$, we have $e \leq 3n + 3g(G) - 6$. ∎

To prove our upper bound, we need to generalize this inequality for the union of several graphs.

THEOREM 8 (generalized Euler's formula). *Let G be a simple graph and let $G_1, \ldots, G_k$ be a partition of G into k parts. Let $n_i = n(G_i) \geq 3$ for each $1 \leq i \leq k$. If every component of each G_i has at least three vertices, then*

$$e \leq 3g(G) + 3\sum_{i=1}^{k}(n_i - 2).$$

Proof. Suppose that the claim is false, and let G together with its partition to graphs $G_1, \ldots, G_k$ be a counterexample that is "smallest" in the following sense:

1. $\sum_{i=1}^{k}(n_i - 2)$ is the smallest possible, and
2. among all graphs that satisfy the first condition, n is the largest possible.

By Proposition 7, we know that $k > 1$. Let us now describe some of the properties of G and its partition:

(i) *Each G_i is connected.* Otherwise, we may assume without loss of generality that G_1 is not connected, i.e., $G_1 = G_1^a \cup G_1^b$, where G_1^a and G_1^b are vertex-disjoint. By the minimality, the partition $G = G_1^a \cup G_1^b \cup G_2 \cup \cdots \cup G_k$ satisfies $e \leq 3g(G) + 3n(G_a) - 6 + 3n(G_b) - 6 + 3\sum_{i=2}^{k}(n_i - 2) < 3g(G) + 3\sum_{i=1}^{k}(n_i - 2)$, which is a contradiction with the fact that G is a counterexample.

(ii) *G is connected.* Otherwise, G is a vertex-disjoint union of two smaller graphs G_a and G_b, and we may assume that $G_a = G_1 \cup \cdots \cup G_t$ and $G_b = G_{t+1} \cup \cdots \cup G_k$ (the graphs G_i are connected; thus they must be subgraphs of one

of these two graphs). By Theorem 3, $g(G) = g(G_a) + g(G_b)$, and since G is a minimal counterexample, we have $e(G_a) \leq 3g(G_a) + 3\sum_{i=1}^{t}(n_i - 2)$ and $e(G_b) \leq 3g(G_b) + 3\sum_{i=t+1}^{k}(n_i - 2)$. Summing these two inequalities results in a contradiction with the fact that G is a counterexample.

(iii) *Each n_i is at least* 4. Otherwise, we may assume that $n_1 = 3$ and let G' be the union of graphs $G_2, \ldots, G_k$. Since $g(G') \leq g(G)$ and G is a minimal counterexample, it follows that $e(G') = e - e(G_1) \leq 3g(G) + 3\sum_{i=2}^{k}(n_i - 2)$. However, $e(G_1) \leq 3 = 3(n_1 - 2)$, and hence $e \leq 3g(G) + 3\sum_{i=1}^{k}(n_i - 2)$, which is contradiction.

(iv) *The minimum degree of each G_i is at least* 3. Otherwise, we may assume that v is a vertex of G_1 with degree $d \leq 2$. Let $G_1' = G_1 - v$, and let G' be the union of graphs $G_1', G_2, G_3, \ldots, G_k$. Suppose that G_1' satisfies the assumptions of the theorem. Since $g(G') \leq g(G)$ and G is a minimal counterexample, we get $e(G') = e(G) - d \leq 3g(G) + 3\sum_{i=1}^{k}(n_i - 2) - 3$, which is again a contradiction.

We need to verify that G_1 satisfies the assumptions of the theorem. This is trivial if v is not a cut-vertex of G_1, since $n_1 \geq 4$ by the previous item. Therefore, if $d = 1$, then the assumptions are satisfied, and we may assume that G_1 does not contain a vertex of degree 1. Let us consider the case that $d = 2$ and v is a cut-vertex. Since $\delta(G_1) \geq 2$, both components of G_1' have at least three vertices; hence in this case G_1' satisfies the assumptions of the theorem as well.

(v) *G is 2-connected.* Otherwise, suppose that $G = G_a \cup G_b$, where G_a and G_b share just a single vertex v. By Theorem 3, $g(G) = g(G_a) + g(G_b)$. Suppose that the graphs $G_1, \ldots, G_t$ are subgraphs of G_a, the graphs $G_{t+1}, \ldots, G_r$ are subgraphs of G_b, and for $r < i \leq k$, $G_i = G_i^a \cup G_i^b$, where G_i^a is a subgraph of G_a and G_i^b is a subgraph of G_b. Since the minimum degree of G_i is at least 3, both G_i^a and G_i^b have at least three vertices. Again, by summing the inequalities $e(G_a) \leq 3g(G_a) + 3\sum_{i=t}^{t}(n_i - 2) + 3\sum_{i=r+1}^{k}(n(G_i^a) - 2)$ and $e(G_b) \leq 3g(G_b) + 3\sum_{i=t+1}^{r}(n_i - 2) + 3\sum_{i=r+1}^{k}(n(G_i^b) - 2)$, we obtain a contradiction with the minimality of G.

(vi) *Each two graphs G_i and G_j share at most one vertex.* Otherwise, if G_{k-1} and G_k share $t \geq 2$ vertices, then let $G_{k-1}' = G_{k-1} \cup G_k$, and apply the theorem on G split into graphs $G_1, \ldots, G_{k-2}, G_{k-1}'$. We obtain $e \leq 3g(G) + 3\sum_{i=1}^{k}(n_i - 2) + 6 - 3t$, which is a contradiction since $6 - 3t \leq 0$.

Let us now fix an embedding of G on a surface of Euler genus $g(G)$. Recall that this embedding is cellular by Theorem 2. Given a vertex v of degree d in G, let $e_0, \ldots, e_{d-1}$ be the edges of G in a cyclic ordering around v. A *segment* is a maximum interval $[a, b]$ such that all the edges $e_a, e_{a+1}, \ldots, e_b$ (with the indices taken modulo d) belong to a single graph G_i. The edges e_a and e_b are called *boundary edges* of the segment. The *length* of the segment is the number of its edges. The embedding of G has the following properties:

- *If a vertex v belongs to at least two parts, then there are at least two segments of edges at v for each of these parts.* Otherwise, suppose that all the edges of G_1 at v form just a single segment. In this case, we may split v into two vertices v_1 and v_2 such that all edges of G_1 at v are incident to v_1 and all the remaining edges at v are incident to v_2. The created graph G' is a counterexample embedded in the same surface with $e(G') = e(G)$ and $n(G') > n(G)$, which is a contradiction to the choice of G.
- *The following configuration ($\star$) of edges cannot appear: $e_1 = vw$ belongs to G_i, all the remaining edges of G_i at v belong to one segment $[a, b]$, and*

the vertex w appears at a face f incident to e_a or e_b. If this were the case, we might redraw G in such a way that e_1 is adjacent to e_a or e_b in the list of edges at v, by drawing it through the face f. We could then again split the vertex v and obtain a contradiction.

We now plug the equality for Δ from Proposition 7 into the formula that we want to prove, thus obtaining the following equivalent inequality:

$$\Delta - 3n + 3\sum_{i=1}^{k} n_i \geq 6k - 6.$$

Therefore, we need to show that either G has long faces or the vertex sets of the graphs G_i have a big overlap. In fact, we prove that if the embedding of the graph G and its partition satisfies all the conditions described above, then the following stronger claim holds:

$$\Delta - 3n + 3\sum_{i=1}^{k} n_i \geq 6k.$$

We proceed by the discharging method. We assign an initial charge to each vertex and each face in the following way: a vertex v that belongs to x of the graphs G_i has initial charge $3(x-1)$. A face of length ℓ has initial charge $\ell - 3$. The sum of these charges is equal to $\Delta - 3n + 3\sum_{i=1}^{k} n_i$.

Next, we move some of this charge to the graphs G_i in such a way that the final charge of each vertex and each face is nonnegative, and the final charge of each G_i is at least 6. Since no charge is lost in the process, the required inequality follows.

We use the following rules to redistribute the charge:

(R1) Each vertex v that belongs to $x \geq 2$ graphs G_i sends charge $3/2$ to each of these graphs.

(R2) Let f be a ≥ 4-face, and let $v_1v_2v_3v_4v_5$ be a subwalk of the facial walk of f such that edges v_2v_3 and v_3v_4 belong to the same graph G_i and neither v_1v_2 nor v_4v_5 belongs to G_i. Then, f sends $1/2$ to G_i through each of v_2 and v_4 (one unit of charge in total).

(R3) Let $f = wv_1v_2w\,v_4v_5$ be a 6-face such that the edges v_1v_2, v_2w, and v_1w belong to a graph G_i and the edges $w\,v_4$, v_4v_5, and v_5w belong to a different graph G_j. Then, f sends $3/2$ to each of G_i and G_j through the vertex w.

(R4) Let f be a face of length at least $t-1$ (where $t > 5$) for which rule (R3) does not apply, and let $v_1v_2 \cdots v_t$ be a subwalk of the facial walk of f such that the edges $v_2v_3, v_3v_4, \ldots,$ and $v_{t-2}v_{t-1}$ belong to the same graph G_i, and neither v_1v_2 nor $v_{t-1}v_t$ belongs to G_i. Then, f sends 1 to G_i through each of v_2 and v_{t-1} (two units of charge in total).

Let us first show that after the rules are applied, the final charge of each vertex and each face is nonnegative. If v is a vertex that belongs to x graphs G_i, then its final charge is zero if $x = 1$ and it is $3(x-1) - 3x/2 = 3x/2 - 3 \geq 0$ if $x \geq 2$ by rule (R1). Now, consider the charge of the faces. Let f be an arbitrary face of G:

(a) If rule (R3) is applied to f, then its final charge is zero.

(b) If f is a 3-face, then either all of its edges belong to the same graph or each of them belongs to a different graph, as otherwise two of the graphs G_i would intersect in at least two vertices. Therefore, no rule applies to f, and the final charge of f is zero.

(c) Finally, suppose that rule (R2) applies a times and rule (R4) applies b times on an ℓ-face f. The final charge of f is $\ell - 3 - a - 2b$; therefore, it suffices to consider the case that $a + 2b + 2 \geq \ell \geq 4$. On the other hand, $\ell \geq 2a + 3b$; hence the final charge is at least $a + b - 3$, and we may assume that $a + b \leq 2$. It follows that $\ell \leq 6$ and exactly two of the graphs G_i contain edges of the face f. Since these two graphs may share only one vertex and the graph is simple, f must be a 6-face consisting of two triangles, $a = 0$ and $b = 2$. But then we obtain case (a), which is covered by rule (R3).

Now, let us consider the charge of the parts. We need to prove that the final charge of each of the parts is at least six. Let G_i be one of the parts, and let Y be the set of vertices that G_i shares with the rest of the graph G. Since G is 2-connected, $|Y| \geq 2$. By rule (R1), the subgraph G_i receives $3|Y|/2$ units of charge, which is at least six if $|Y| \geq 4$. Therefore, it suffices to consider the cases $|Y| = 2$ and $|Y| = 3$.

We call a boundary edge e of a segment of G_i at a vertex $v \in Y$ *rich* if e does not connect v with another vertex of Y. Let $e = vw$ be a rich edge, and let f_e be a face that contains e and an edge incident to v that does not belong to G_i. Since $w \notin Y$, all the edges incident to w must belong to G_i; hence one of rules (R2), (R3), or (R4) applies and f_e sends at least $1/2$ units of charge through v to G_i.

Suppose first that $|Y| = 3$. Let v be an arbitrary vertex in Y. The edges of G_i at v form at least two segments. By property (iv), the degree of v in G_i is at least 3; hence there are at least three boundary edges incident with v. Since $|Y \backslash \{v\}| = 2$, at least one of these edges is rich; hence G_i receives at least $1/2$ units of charge through v. Therefore, G_i receives $9/2$ units of charge by rule (R1) and at least $1/2$ units of charge by rules (R2)–(R4) through each vertex of Y, which sum to at least six units of charge.

Suppose now that $|Y| = 2$. The graph G_i receives three units of charge by rule (R1). We prove that at least $3/2$ units of charge are sent to G_i through each vertex of Y by rules (R2)–(R4), thus showing that G_i receives at least six units of charge. Suppose for the sake of contradiction that less than $3/2$ units of charge are sent to G_i through a vertex $v \in Y$. Then, there are at most two rich edges incident with v. On the other hand, G_i has at least two segments at v, the degree of v is at least 3 by property (iv), and $Y \backslash \{v\}$ consists of only one vertex w; thus at least two rich edges are incident with v. Hence, we conclude that there are exactly two rich edges at v. This is possible only in the following cases:

- *The degree of v in G_i is three, and each of the edges of G_i incident with v forms a segment of length one.* However, note that in this case, each of the four (not necessarily distinct) faces incident with the rich edges sends $1/2$ units of charge through v, for a total of two units of charge.
- *There are exactly two segments of G_i at v and one of them is of length one.* Let $e_0 = vu_0$ be the edge of the segment of length one, and let $e_1 = vu_1$ and $e_2 = vu_2$ be the boundary edges of the other segment. Note that $u_1 \neq u_2$, as the degree of v is at least 3. If $w \neq u_0$ (say, $w = e_1$), then each of the faces incident with e_0 send $1/2$ units of charge through v and the face f_{e_2} sends $1/2$ units of charge, for a total of $3/2$ units.

  Let us now consider the case that $w = u_0$. The graph G_i receives $1/2$ units of charge through v for each of e_1 and e_2. If rules (R3) or (R4) were applied at v at least once, G_i would receive additional $1/2$ units of charge, contradicting the choice of v. Let us assume that this is not the case. Let w_1 and w_2 be the vertices following u_1 and u_2 in the facial walks of f_{e_1} and f_{e_2}, respectively. For $i = 1, 2$, the vertices w_i and v are both neighbors of u_i;

hence $w_1 \neq v \neq w_2$. The edges following w_1 and w_2 in the facial walks do not belong to G_i, since otherwise one of rules (R3) or (R4) applies. This means that $w_1, w_2 \in Y$, and hence $w_1 = w_2 = w$. This is the forbidden configuration ($\star$); hence we obtain a contradiction with the assumption that less than $3/2$ is sent through the vertex v.

It follows that the final charge of each of the graphs G_i is at least six, and thus we conclude that $\Delta + 3(\sum_{i=1}^{k} n_i - n) \geq 6k$, which finishes the proof. □

Theorem 8 is tight; for example, the equality is obtained for a disjoint union of k triangulations, or graphs are obtained from this graph by identifying the vertices in such a way that all edges of each graph form one segment at each vertex. Also, it is not possible to relax the condition on the number of vertices in G_i, as the claim is false if each G_i is just an edge.

3. Upper bound. We are now ready to prove the upper bound on the k-chromatic number $\chi_k(G)$ of a graph G of Euler genus g. Our method is similar to the one used by Furedi et al. [5], except that we use a better estimate on the number of edges of G obtained from Theorem 8.

Proof of Theorem 1. Let us embed G in a surface of Euler genus g. Let $G_1, \ldots, G_k$ be a partition of G into k parts. For each i, let $G'_i \subseteq G_i$ be a critical subgraph of G_i such that $\chi(G'_i) = \chi(G_i) = c_i$. We may assume that $c_1 \geq c_2 \geq \cdots \geq c_k$. Let t be the largest number such that $c_t \geq 7$. Thus, $t = 0$ if $c_1 \leq 6$. We bound the sum of chromatic numbers of the graphs $G_1, \ldots, G_t$. Let $n_i = n(G'_i)$.

Let $G' = G'_1 \cup \cdots \cup G'_t$ and $e' = e(G')$. Using Theorem 8, we get

$$2e' \leq 6g + 6\sum_{i=1}^{t}(n_i - 2).$$

On the other hand, the minimum degree of each G'_i is at least $c_i - 1$; hence $(c_i - 1)n_i \leq 2e(G'_i)$. This implies that

$$\sum_{i=1}^{t}(c_i - 1)n_i \leq 6g + 6\sum_{i=1}^{t}(n_i - 2).$$

Using the face that $c_i \geq 7$ and $n_i \geq c_i$, we obtain

$$\sum_{i=1}^{t}[(c_i - 7/2)^2 - 49/4] = \sum_{i=1}^{t}(c_i - 7)c_i \leq \sum_{i=1}^{t}(c_i - 7)n_i \leq 6g - 12t.$$

By the inequality between the arithmetic and quadratic means,

$$\frac{1}{t}\left[\sum_{i=1}^{t}(c_i - 7/2)\right]^2 \leq 6g + t/4,$$

from which we infer

$$\sum_{i=1}^{t} c_i \leq \frac{7t + \sqrt{24tg + t^2}}{2}.$$

Taking into account the graphs $G_{t+1}, \ldots, G_k$, we get

$$\sum_{i=1}^{k} c_i \leq \frac{7t + \sqrt{24tg + t^2}}{2} + 6(k - t).$$

If $t \leq g$, this expression is increasing in t; thus we obtain

$$\sum_{i=1}^{k} c_i \leq \frac{7k + \sqrt{24kg + k^2}}{2}.$$

Since the expression on the left-hand side is an integer, we may round the expression on the right-hand side down, thus finishing the proof of this theorem. ∎

4. Lower bound. The proof of the upper bound hints at how the lower bound examples should look. For each of the graphs in the partition, we should have $c_i = n_i$; hence all the graphs G_i should be complete. Also, since we used the inequality between the arithmetic and quadratic means, their sizes should be the same. This is possible only for special values of g and k. For example, consider the case $g = \frac{1}{6}k(t-3)(t-4)$ for some $t \geq 4$, $t \equiv 0, 1 \pmod 3$. Then, K_t can be embedded in a surface of genus g/k (K_7 cannot be embedded in the Klein bottle, but it can be embedded in the torus), according to Theorem 4. By Theorem 3, the disjoint union of k complete graphs on t vertices can be embedded in a surface S of genus g; hence

$$\chi_k(S) \geq kt = \frac{7k + \sqrt{24kg + k^2}}{2}.$$

For general g and k, we cannot hope for a nice formula like the one in Theorem 1; thus we would be satisfied with some description of the best possible example. A natural guess is that this example is a disjoint union of cliques. We were not able to prove that this is the case. The best result that we obtained in this direction is the following proposition.

PROPOSITION 9. *Let $G_1, \ldots, G_k$ be a partition of a graph G of Euler genus g into k parts, and let $c_i = \chi(G_i) \geq 7$ for each i. Let G_i' be a c_i-critical subgraph of G_i. Suppose that $c_i \equiv 0, 1 \pmod 3$ whenever G_i' is a clique. Then, the disjoint union of the cliques $K_{c_1}, \ldots, K_{c_k}$ has Euler genus at most g.*

Proof. Let $e' = e(G_1' \cup \cdots \cup G_k')$ and $n_i = n(G_i')$, and let $\delta_i = 0$ if G_i' is a clique and $\delta_i = c_i - 3$ otherwise. By Theorem 8,

$$2e' \leq 6g + 6\sum_{i=1}^{k}(n_i - 2).$$

On the other hand, using Theorem 6, we get

$$2e' \geq \sum_{i=1}^{k}(c_i - 1)n_i + \delta_i.$$

Therefore, we obtain

$$\begin{aligned} g &\geq \frac{1}{6}\sum_{i=1}^{k}(c_i - 7)n_i + 12 + \delta_i \\ &\geq \sum_{i=1}^{k}\frac{1}{6}((c_i - 7)c_i + 12 + \delta_i) \\ &\geq \sum_{i=1}^{k}\left\lceil \frac{1}{6}(c_i - 3)(c_i - 4)\right\rceil = \sum_{i=1}^{k} g(K_{c_i}), \end{aligned}$$

where the last inequality holds because $\delta_i \geq 4$ whenever $c_i \equiv 2 \pmod 3$, by the assumptions of the lemma. The statement of the lemma follows from Theorem 3. ∎

5. Conclusions. Let us call the complete graph K_n *bad* if it does not triangulate the minimal surface in which it can be embedded, i.e., $n \equiv 2 \pmod 3$. Proposition 9 shows that the best values of χ_k are achieved for disjoint unions of cliques, unless bad cliques appear in the partition. It is natural to ask whether the restriction on the appearance of the bad cliques is necessary, or whether it is always possible to "disentangle" cliques.

PROBLEM 1. *Let $G_1, \ldots, G_k$ be a partition of a graph G into k parts such that each subgraph G_i is a clique. Is it true that the vertex-disjoint union of the cliques G_i can be embedded in a surface of Euler genus $g(G)$?*

For $k = 2$, this follows from Theorem 3. The proof of Theorem 8 shows that unless the graphs in the partition can be trivially disentangled, we may decrease the bound by 6, which implies that the answer to Problem 1 is positive for $k = 3$.

One way to answer the question in Problem 1 positively for $k \geq 4$ would be to improve Theorem 8 by decreasing the right-hand side of the inequality by 2 for each bad clique in the partition. Another way is suggested by the following conjecture of Stiebitz and Škrekovski [14].

CONJECTURE 1. *Let G be an edge-disjoint union of a clique K and an arbitrary graph H. Let H' be the graph obtained from H by contracting the set $V(K)$ to a single vertex. Then, $g(H') + g(K) \leq g(G)$.*

Because two complete graphs in a partition of a graph into k parts cannot share more than one vertex, it is easy to show by induction that Conjecture 1 implies a positive answer to Problem 1.

In our considerations, we do not distinguish between orientable and nonorientable surfaces; we focus only on their Euler genus. While asymptotically there does not seem to be much difference, for some values k and g the results may differ.

We have provided (almost) matching upper and lower bounds for a k-chromatic number of graphs with bounded genus g, assuming that the genus is large enough regarding k. The reason our techniques cannot be directly applied in the case in which k is larger than g is that we would need to consider critical graphs with chromatic number ≤ 6. Graphs with chromatic number ≤ 4 are easy to handle; we may assume that they appear only as K_4-disjoint with the rest of the graph, since they are planar and hence do not affect the genus of the graph. However, graphs with chromatic number 5 and 6 are difficult to deal with. For chromatic number 6, the list of critical graphs is known only for surfaces with $g \leq 2$, and for the chromatic number 5, there even are infinitely many of them on each surface with $g \geq 1$. Nevertheless, it might be interesting to determine the exact values of χ_k for some special cases, e.g., for graphs embedded in the torus or in the projective plane.

REFERENCES

[1] K. APPEL AND W. HAKEN, *Every Planar Map is Four Colorable*, Contemp. Math., 98, AMS, Providence, RI, 1977.

[2] J. BATTLE, F. HARARY, Y. KODOMA, AND J. W. T. YOUNGS, *Additivity of the genus of a graph*, Bull. Amer. Math. Soc., 68 (1962), pp. 565–568.

[3] G. A. DIRAC, *A theorem of R. L. Brooks and a conjecture of H. Hadwiger*, Proc. London Math. Soc. (3), 7 (1957), pp. 161–195.

[4] P. FRANKLIN, *A six colour problem*, J. Math. Phys., 13 (1934), pp. 363–369.

[5] Z. FÜREDI, A. V. KOSTOCHKA, M. STIEBITZ, R. ŠKREKOVSKI, AND D. B. WEST, *Nordhaus–Gaddum-type theorems for decompositions into many parts*, J. Graph Theory, 50 (2005), pp. 273–292.

[6] P. J. HEAWOOD, *Map colour theorem*, Quart. J. Pure Appl. Math., 24 (1890), pp. 332–338.

[7] B. Mohar and C. Thomassen, *Graphs on Surfaces*, Johns Hopkins University Press, Baltimore, MD, 2001.
[8] E. A. Nordhaus and J. W. Gaddum, *On complementary graphs*, Amer. Math. Monthly, 63 (1956), pp. 175–177.
[9] J. Plesník, *Bounds on the chromatic numbers of multiple factors of a complete graph*, J. Graph Theory, 2 (1978), pp. 9–17.
[10] G. Ringel and J. W. T. Youngs, *Solution of the Heawood map-coloring problem*, Proc. Nat. Acad. Sci. U.S.A., 60 (1968), pp. 438–445.
[11] G. Ringel, *Map Color Theorem*, Springer-Verlag, New York, 1974.
[12] N. Robertson, D. P. Sanders, P. Seymour, and R. Thomas, *The four colour theorem*, J. Combin. Theory Ser. B, 70 (1997), pp. 2–44.
[13] S. Stahl and L. W. Beineke, *Blocks and the non-orientable genus of graphs*, J. Graph Theory, 1 (1977), pp. 75–78.
[14] M. Stiebitz and R. Škrekovski, *A map colour theorem for the union of graphs*, J. Combin. Theory Ser. B, 96 (2006), pp. 20–37.
[15] T. Watkinson, *A theorem of the Nordhaus–Gaddum class*, Ars Combin., 20-B (1985), pp. 35–42.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 487–510

ALGORITHMS FOR DUALIZATION OVER PRODUCTS OF PARTIALLY ORDERED SETS*

KHALED M. ELBASSIONI†

Abstract. Let $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$ be the product of n partially ordered sets (posets). Given a subset $\mathcal{A} \subseteq \mathcal{P}$, we consider problem DUAL$(\mathcal{P}, \mathcal{A}, \mathcal{B})$ of extending a given partial list $\mathcal{B}$ of maximal independent elements of $\mathcal{A}$ in $\mathcal{P}$. We give quasi-polynomial time algorithms for solving problem DUAL$(\mathcal{P}, \mathcal{A}, \mathcal{B})$ when each poset $\mathcal{P}_i$ belongs to one of the following classes: (i) semilattices of bounded width, (ii) forests, that is, posets with acyclic underlying graphs, with either bounded in-degrees or out-degrees, or (iii) lattices defined by a set of real closed intervals.

Key words. enumeration algorithms, forests, hypergraph transversals, infrequent elements, lattices, monotone properties, monotone generation, ordered sets, duality testing

AMS subject classifications. 68Q25, 68R01

DOI. 10.1137/050622250

1. Introduction.

Let $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$ be the product of n partially ordered sets (posets). Denote by $\preceq$ the precedence relation in $\mathcal{P}$ and also in $\mathcal{P}_1, \ldots, \mathcal{P}_n$, i.e., if $p = (p_1, \ldots, p_n) \in \mathcal{P}$ and $q = (q_1, \ldots, q_n) \in \mathcal{P}$, then $p \preceq q$ in $\mathcal{P}$ if and only if $p_1 \preceq q_1$ in $\mathcal{P}_1$, $p_2 \preceq q_2$ in $\mathcal{P}_2, \ldots$, and $p_n \preceq q_n$ in $\mathcal{P}_n$. For a set $\mathcal{A} \subseteq \mathcal{P}$, let $\mathcal{A}^+ = \{x \in \mathcal{P} \mid x \succeq a \text{ for some } a \in \mathcal{A}\}$ and $\mathcal{A}^- = \{x \in \mathcal{P} \mid x \preceq a \text{ for some } a \in \mathcal{A}\}$ denote, respectively, the ideal and filter generated by $\mathcal{A}$. For simplicity, we shall use p^+ and p^- to denote $\{p\}^+$ and $\{p\}^-$ for any $p \in \mathcal{P}$. Any element in $\mathcal{P} \setminus \mathcal{A}^+$ is called *independent of* $\mathcal{A}$. Let $\mathcal{I}(\mathcal{A})$ be the set of all maximal independent elements for $\mathcal{A}$, also called the *dual* of $\mathcal{A}$ in $\mathcal{P}$:

$$\mathcal{I}(\mathcal{A}) \stackrel{\text{def}}{=} \{p \in \mathcal{P} \mid p \notin \mathcal{A}^+ \text{ and } (q \in \mathcal{P},\ q \succeq p,\ q \neq p \ \Rightarrow\ q \in \mathcal{A}^+)\}.$$

Then we have the following decomposition of $\mathcal{P}$:

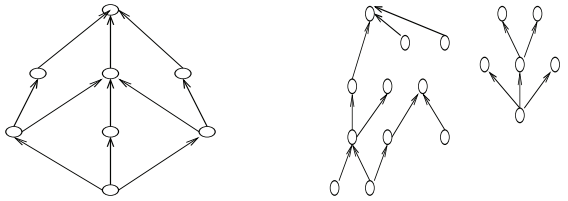
$$\mathcal{A}^+ \cap \mathcal{I}(\mathcal{A})^- = \emptyset, \quad \mathcal{A}^+ \cup \mathcal{I}(\mathcal{A})^- = \mathcal{P}. \tag{1}$$

Call $\mathcal{A}$ an *antichain* if no two elements are comparable in $\mathcal{P}$. In this paper, we are concerned with the following *dualization* problem:

DUAL$(\mathcal{P}, \mathcal{A}, \mathcal{B})$**:** *Given an antichain $\mathcal{A} \subseteq \mathcal{P}$ in a poset $\mathcal{P}$ and a collection of maximal independent elements $\mathcal{B} \subseteq \mathcal{I}(\mathcal{A})$, either find a new maximal independent element $x \in \mathcal{I}(\mathcal{A}) \setminus \mathcal{B}$, or state that the given collection is complete: $\mathcal{B} = \mathcal{I}(\mathcal{A})$.*

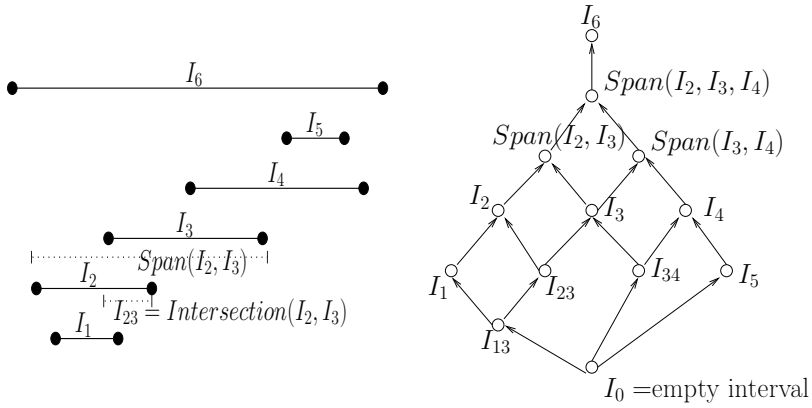
If $\mathcal{P}$ is the Boolean cube, i.e., $\mathcal{P}_i = \{0, 1\}$ for all $i = 1, \ldots, n$, then the above dualization problem reduces to the well-known *hypergraph transversal* problem, which calls for enumerating all minimal subsets $X \subseteq V$ that intersect all edges of a given hypergraph $\mathcal{H} \subseteq 2^V$. The complexity of this problem is still an important open question, for which the currently best known algorithm [FK96] runs in quasi-polynomial time $poly(n, m) + m^{o(\log m)}$, where $m = |\mathcal{A}| + |\mathcal{B}|$, providing strong evidence that the problem is unlikely to be NP-hard. More generally, when each $\mathcal{P}_i$ is a chain, that is,

*Received by the editors January 8, 2005; accepted for publication (in revised form) August 10, 2008; published electronically February 4, 2009. A preliminary version of this paper, containing some of the results, appeared in Elbassioni [Elb02b, Elb02a, Elb06].
http://www.siam.org/journals/sidma/23-1/62225.html
†Max-Planck-Institut für Informatik, 66123 Saarbrücken, Germany (elbassio@mpi-sb.mpg.de).



(a) A lattice with $W = 3$. (b) A forest with $d = 3$.

FIG. 1. *Lattices and forests.*



(a) A set of intervals $\mathbb{I}_1$. (b) The corresponding lattice of intervals $\mathcal{L}_1$.

FIG. 2. *The lattice of intervals.*

a totally ordered set, the problem was considered in [BEG$^+$02], where it was shown that the algorithms of [FK96] can be extended to work in quasi-polynomial time, regardless of each chain's size. It is natural to investigate whether these results can be extended further to wider classes of partially ordered sets. In this paper, we achieve this for the cases when each $\mathcal{P}_i$ is the following:

(i) a join (or meet) semilattice with bounded width (see Figure 1(a)),
(ii) a forest, that is, a poset in which the underlying undirected graph of the precedence graph is acyclic (see Figure 1(b)), and in which either the in-degree or the out-degree of each element is bounded, and
(iii) the lattice of intervals defined by a set of intervals on the real line $\mathbb{R}$ (see Figure 2): Let $\mathbb{I}_i$ be a set of intervals in $\mathbb{R}$, and let $\mathcal{L}_i$ be the *lattice of intervals* whose elements are all possible intersections and spans defined by the intervals in $\mathbb{I}_i$ and ordered by containment. The meet of any two intervals in $\mathcal{L}_i$ is their *intersection*, and the join is their *span*, i.e., the minimum interval containing both of them.

We remark that for case (i), all posets $\mathcal{P}_i$ must be of the same type: either all posets are join semilattices, or all of them are meet semilattices. Without loss of generality, we will consider only join semilattices.

1.1. Main results. Here is a more formal description of the results in this paper. For $x \in \mathcal{P}_i$, denote by $x^\perp$ the set of immediate predecessors of x, i.e.,

$$x^\perp = \{y \in \mathcal{P}_i \mid y \prec x,\ (\nexists z \in \mathcal{P}_i : y \prec z \prec x)\},$$

and let in-deg$(\mathcal{P}_i) = \max\{|x^{\perp}| : x \in \mathcal{P}_i\}$. Similarly, denote by $x^{\top}$ the set of immediate successors of x, and let out-deg$(\mathcal{P}_i) = \max\{|x^{\top}| : x \in \mathcal{P}_i\}$. Throughout this paper, we shall use the notation $m \stackrel{\text{def}}{=} |\mathcal{A}| + |\mathcal{B}|$, $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$, $d \stackrel{\text{def}}{=} \max_{i\in[n]} \min\{\text{in-deg}(\mathcal{P}_i), \text{out-deg}(\mathcal{P}_i)\}$, and $\mu = \mu(\mathcal{P}) \stackrel{\text{def}}{=} \max\{|\mathcal{P}_i| : i \in [n]\}$. Finally, denote by $W(\mathcal{P}_i)$ the *width* of $\mathcal{P}_i$, i.e., the maximum size of an antichain in $\mathcal{P}_i$, let $W = W(\mathcal{P}) \stackrel{\text{def}}{=} \max_{i\in[n]}\{W(\mathcal{P}_i)\}$ be the maximum width of the n posets, and write $\gamma(W) \stackrel{\text{def}}{=} 2W^2 \ln(W+1)$. A join (respectively, meet) semi-lattice is a poset $\mathcal{P}$ in which every two elements $x, y \in \mathcal{P}$ have a unique minimum upper-bound, called the *join* $x \vee y$ (respectively, a unique maximum lower-bound, called the *meet* $x \wedge y$).

THEOREM 1. *Problem DUAL$(\mathcal{L}, \mathcal{A}, \mathcal{B})$ can be solved in poly$(n, \mu(\mathcal{L})) \cdot m^{\gamma(W(\mathcal{L}))\cdot o(\log m)}$ time if $\mathcal{L}$ is a product of join semilattices.*

THEOREM 2. *Problem DUAL$(\mathcal{P}, \mathcal{A}, \mathcal{B})$ can be solved in poly$(n, \mu(\mathcal{P})) \cdot m^{d\cdot o(\log m)}$ time if $\mathcal{P}$ is a product of forests.*

THEOREM 3. *Problem DUAL$(\mathcal{L}, \mathcal{A}, \mathcal{B})$ can be solved in $k^{O(\log^2 k)}$ time if $\mathcal{L}$ is a product of lattices of intervals, where $k = |\mathcal{A}| + |\mathcal{B}| + \sum_{i=1}^{n} |\mathcal{L}_i|$.*

Note that Theorem 3 strengthens Theorem 1 for the special case of the product of lattices of intervals. Indeed, for the lattice of intervals $\mathcal{L}_i$, defined by a set of intervals $\mathbb{I}_i$, we have $W(\mathcal{L}_i) = O(|\mathbb{I}_i|)$ and $|\mathcal{L}_i| = O(|\mathbb{I}_i|^2)$, and these bounds are tight. Thus, for this special case, the result of Theorem 1 gives an exponential algorithm in the maximum number of intervals $\max_{i=1}^{n} |\mathbb{I}_i|$, while Theorem 3 gives a quasi-polynomial bound.

In the next section, we consider some applications that motivate our consideration of problem DUAL$(\mathcal{P}, \mathcal{A}, \mathcal{B})$. In section 3, we describe the general approach we use for solving the dualization problem. We prove Theorems 1, 2, and 3 in sections 4, 5, and 6, respectively.

2. Some applications. Let $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$ be a partially ordered set. Consider a monotone property $\pi : \mathcal{P} \mapsto \{0, 1\}$ defined over the elements of $\mathcal{P}$: if $x \in \mathcal{P}$ satisfies π, i.e., $\pi(x) = 1$, then any $y \succeq x$ satisfies π. We assume that π is described by a polynomial-time *satisfiability oracle* $\mathcal{O}_\pi$, i.e., an algorithm that can decide whether a given vector $x \in \mathcal{P}$ satisfies π, in time polynomial in n and the size $|\pi|$ of the input description of π. Denote by $\mathcal{F}_\pi$ and $\mathcal{G}_\pi$, respectively, the families of minimal elements satisfying property π, and maximal elements not satisfying property π. Then it is clear that $\mathcal{G}_\pi = \mathcal{I}(\mathcal{F}_\pi)$. Given a monotone property π, we consider the problem of jointly generating the families $\mathcal{F}_\pi$ and $\mathcal{G}_\pi$:

***GEN*$(\mathcal{P}, \mathcal{F}_\pi, \mathcal{G}_\pi, \mathcal{X}, \mathcal{Y})$:** *Given a monotone property π, represented by a satisfiability oracle $\mathcal{O}_\pi$, and two explicitly listed vector families $\mathcal{X} \subseteq \mathcal{F}_\pi \subseteq \mathcal{P}$ and $\mathcal{Y} \subseteq \mathcal{G}_\pi \subseteq \mathcal{P}$, either find a new element in $(\mathcal{F}_\pi \setminus \mathcal{X}) \cup (\mathcal{G}_\pi \setminus \mathcal{Y})$, or state that these families are complete: $(\mathcal{X}, \mathcal{Y}) = (\mathcal{F}_\pi, \mathcal{G}_\pi)$.*

For a given monotone property π, described by a satisfiability oracle $\mathcal{O}_\pi$, we can generate both $\mathcal{F}_\pi$ and $\mathcal{G}_\pi$ simultaneously by starting with $\mathcal{X} = \mathcal{Y} = \emptyset$ and solving problem GEN$(\mathcal{P}, \mathcal{F}_\pi, \mathcal{G}_\pi, \mathcal{X}, \mathcal{Y})$ for a total of $|\mathcal{F}_\pi| + |\mathcal{G}_\pi| + 1$ times, incrementing in each iteration either $\mathcal{X}$ or $\mathcal{Y}$ by the newly found vector $x \in (\mathcal{F}_\pi \setminus \mathcal{X}) \cup (\mathcal{G}_\pi \setminus \mathcal{Y})$, according to the answer of the oracle $\mathcal{O}_\pi$, until we have $(\mathcal{X}, \mathcal{Y}) = (\mathcal{F}_\pi, \mathcal{G}_\pi)$.

The following result, relating the time complexity of joint generation to that of dualization, is a straightforward generalization of a similar result known for the binary case [BI95, GK99].

PROPOSITION 1. *Problem GEN$(\mathcal{P}, \mathcal{F}_\pi, \mathcal{I}(\mathcal{F}_\pi), \mathcal{X}, \mathcal{Y})$ can be solved in time $O\left(\sum_{i=1}^{n} |\mathcal{P}_i|\right)\ (n|\mathcal{X}||\mathcal{Y}| + T_\pi) + T_{dual}$ for any monotone property π defined by a*

satisfiability oracle $\mathcal{O}_\pi$, where T_π is the worst-case running time of the oracle on any $x \in \mathcal{P}$, and T_{dual} denotes the time required to solve problem DUAL$(\cdot,\cdot,\cdot)$.

When one the two families, say $\mathcal{I}(\mathcal{F}_\pi)$, is bounded polynomially (or quasi-polynomially) in size by the other, i.e.,

$$|\mathcal{I}(\mathcal{F}_\pi)| \leq \text{poly}(|\pi|, |\mathcal{F}_\pi|), \tag{2}$$

then it follows from the above proposition that all of the elements of the family $\mathcal{F}_\pi$ can be generated in *total quasi-polynomial time* quasi-poly$(|\pi|, |\mathcal{F}_\pi|)$.

Problem GEN$(\mathcal{P}, \mathcal{F}_\pi, \mathcal{I}(\mathcal{F}_\pi), \mathcal{X}, \mathcal{Y})$ arises in several applications and in a variety of fields, including artificial intelligence [EG95], game theory [Gur75], reliability theory [BEGK04, Col87], database theory [BGKM03, EG95, GMKT97], integer programming [BEG$^+$02, KBE$^+$07, LLK80], learning theory [AB92], and data mining [AIS93, KBE$^+$07, BGKM03]. In the next subsections we consider three such applications.

2.1. Monotone systems of linear inequalities. Let $A \in \mathbb{R}^{r\times n}$ be a given nonnegative real matrix, $b \in \mathbb{R}^r$ be a given r-vector, and $c \in \mathbb{R}^n_+$ be a given nonnegative n-vector, and consider the system of linear inequalities

$$Ax \geq b, \quad x \in \mathcal{C} = \{x \in \mathbb{Z}^n \mid 0 \leq x \leq c\}. \tag{3}$$

For $x \in \mathcal{C}$, let $\pi(x)$ be the property that x satisfies (3). Then the families $\mathcal{F}_\pi$ and $\mathcal{G}_\pi$ correspond, respectively, to the minimal feasible and maximal infeasible vectors for (3). Proposition 1 implies that problem GEN$(\mathcal{C}, \mathcal{F}_\pi, \mathcal{I}(\mathcal{F}_\pi), \mathcal{X}, \mathcal{Y})$ is polynomially equivalent to dualization over the chain product $\mathcal{C}$. Furthermore, it was shown in [BEG$^+$02] that an inequality of the form (2) holds, namely, $|\mathcal{I}(\mathcal{F}_\pi)| \leq rn|\mathcal{F}_\pi|$. Thus, all minimal feasible solutions for (3) can be generated in quasi-polynomial time (see [BEG$^+$02] for more details).

2.2. Maximal frequent and minimal infrequent elements in products of posets. Let $\mathcal{P} \stackrel{\text{def}}{=} \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$ be the product of n explicitly given posets. Consider a database $\mathcal{D} \subseteq \mathcal{P}$ of transactions, each of which is an n-dimensional vector of attribute values over $\mathcal{P}$. For an element $p \in \mathcal{P}$, let us denote by

$$S(p) = S_{\mathcal{D}}(p) \stackrel{\text{def}}{=} \{q \in \mathcal{D} \mid q \succeq p\}$$

the set of transactions in $\mathcal{D}$ that *support* $p \in \mathcal{P}$. Note that, by this definition, the function $|S(\cdot)| : \mathcal{P} \mapsto \{0, 1, \ldots, |\mathcal{D}|\}$ is an antimonotone function, i.e., $|S(p)| \leq |S(q)|$, whenever $p \succeq q$. Given $\mathcal{D} \subseteq \mathcal{P}$ and an integer threshold t, let us say that an element $p \in \mathcal{P}$ is t-frequent if it is supported by at least t transactions in the database, i.e., if $|S_{\mathcal{D}}(p)| \geq t$. Conversely, $p \in \mathcal{P}$ is said to be t-infrequent if $|S_{\mathcal{D}}(p)| < t$. For each $x \in \mathcal{P}$, let $\pi(x)$ be the property that x is t-infrequent. Then π is a monotone property and the families $\mathcal{F}_\pi$ and $\mathcal{G}_\pi$ correspond, respectively, to minimal t-infrequent and maximal t-frequent elements for $\mathcal{D}$.

The joint generation of maximal frequent and minimal infrequent elements of a database can be used for finding the so-called *association rules* in data mining applications [AIS93]. If the database $\mathcal{D}$ contains categorical (e.g., zip code, make of car) or quantitative (e.g., age, income) attributes, and the corresponding posets $\mathcal{P}_i$ are total orders, then the above generation problems can be used to mine the so-called *quantitative* association rules [SA96]. More generally, each attribute a_i in the database can assume values belonging to some partially ordered set $\mathcal{P}_i$. For example, [SA95]

describes applications where items in the database belong to sets of *taxonomies* (or *is-a hierarchies*), and proposes several algorithms for mining association rules among these hierarchical data (see also [HCC93, HF95]). Proposition 1 and Theorems 1 and 2 imply that, for databases $\mathcal{D} \subseteq \mathcal{P}$ where the underlying precedence graph of each poset $\mathcal{P}_i$ is a rooted tree (is-a hierarchy), or where each poset $\mathcal{P}_i$ is a join semilattice of bounded width, and for any integer threshold t, all maximal frequent elements and all minimal infrequent elements can be jointly generated in quasi-polynomial time (the binary case was considered in [BGKM03]).

2.3. Sparse boxes for multidimensional data. Let $\mathcal{S}$ be a set of points in $\mathbb{R}^n$, and $k \leq |\mathcal{S}|$ be a given integer. A *maximal k-box* is a closed n-dimensional rectangle which contains at most k points of $\mathcal{S}$ in its interior, and which is maximal with respect to this property (i.e., cannot be extended in any direction without strictly enclosing more points of $\mathcal{S}$). Suppose we are interested in generating the family $\mathcal{F}_{\mathcal{S},k}$ of maximal k-boxes, defined by the set of points $\mathcal{S}$. Then, without any loss of generality, we may consider the generation of maximal k-boxes contained in a fixed bounded box D containing all points of $\mathcal{S}$ in its interior. Let us further note that the ith coordinate of each vertex of such a box is the same as p_i for some $p \in \mathcal{S}$, or the ith coordinate of a vertex of D; hence all of these coordinates belong to a finite set of cardinality at most $|\mathcal{S}| + 2$. In other words, we can view $\mathcal{F}_{\mathcal{S},k}$ as a set of boxes with vertices belonging to such a finite grid.

For $i = 1, \ldots, n$, consider the set of projection points $\mathcal{S}_i \stackrel{\text{def}}{=} \{p_i \mid p \in \mathcal{S}\}$, and let $\mathcal{L}_i$ be the *lattice of intervals* whose elements are the different intervals defined by the projection points $\mathcal{S}_i$ and ordered by containment. The minimum element l_i of $\mathcal{L}_i$ corresponds to the empty interval I_0. A 2-dimensional example is shown in Figure 3. Let $\mathcal{L} = \mathcal{L}_1 \times \cdots \times \mathcal{L}_n$; then each element x of $\mathcal{L}$, with $x_i \neq l_i$ for all $i \in [n]$, represents a box containing some points of $\mathcal{S}$. For $x \in \mathcal{L}$, let $\pi(x)$ be the property that the box defined by x contains at least $k+1$ points of S in its interior. Then the sets $\mathcal{G}_\pi$ and $\mathcal{F}_\pi$ can be identified, respectively, with the set of maximal k-boxes and the set of minimal boxes of $x \in \mathcal{L}$ which contain at least $k+1$ points of $\mathcal{S}$ in their interior. Furthermore, it can be shown [KBE+07] that $|\mathcal{F}_\pi| \leq |\mathcal{S}||\mathcal{G}_\pi|$. Thus, Proposition 1 and Theorem 3 imply that the family $\mathcal{F}_{\mathcal{S},k}$ can be generated in quasi-polynomial time (see [KBE+07] for more details).

The problem of generating all of the elements of $\mathcal{F}_{\mathcal{S},0}$ has been studied in the machine learning and computational geometry literatures (see [CDL86, EGLM03, Orl90]) and is motivated by the discovery of missing associations or "holes" in data mining applications (see [AMS+96, LKH97, BLQ98]).

3. General approach.

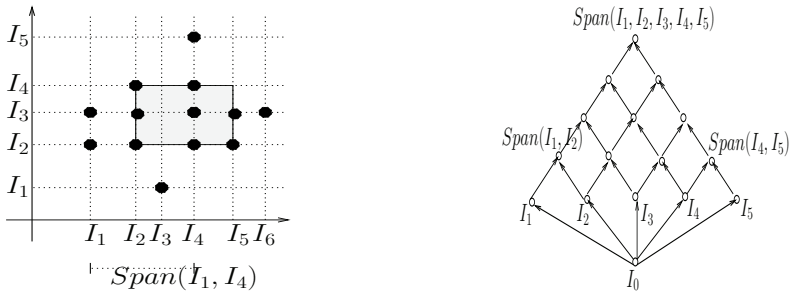
3.1. Preliminaries. Given two subsets $\mathcal{A} \subseteq \mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$ and $\mathcal{B} \subseteq \mathcal{I}(\mathcal{A})$, we say that $\mathcal{B}$ is *dual to* $\mathcal{A}$ if $\mathcal{B} = \mathcal{I}(\mathcal{A})$. By (1), this condition is equivalent to $\mathcal{A}^+ \cup \mathcal{B}^- = \mathcal{P}$, and

$$a \not\preceq b \quad \forall a \in \mathcal{A}, b \in \mathcal{B}. \tag{4}$$

Thus problem DUAL$(\mathcal{P}, \mathcal{A}, \mathcal{B})$ can be equivalently stated as follows:

DUAL$(\mathcal{P}, \mathcal{A}, \mathcal{B})$**:** *Given antichains $\mathcal{A}, \mathcal{B} \subseteq \mathcal{P}$ satisfying* (4)*, check if there an $x \in \mathcal{P} \setminus (\mathcal{A}^+ \cup \mathcal{B}^-)$.*

Having found a *solution* to DUAL$(\mathcal{P}, \mathcal{A}, \mathcal{B})$, i.e., an element $x \in \mathcal{P} \setminus (\mathcal{A}^+ \cup \mathcal{B}^-)$, it can be extended to a maximal element x^* with the same property in $O(n\mu|\mathcal{A}|)$



(a) A 2-dimensional pointset and a maximal 1-box. (b) The lattice of intervals $\mathcal{L}_1$.

FIG. 3. *Maximal sparse boxes: the shaded box has at most $t = 1$ point in its interior.*

time. This can be done by initializing $c(a) = |\{i \in [n] : a_i \not\preceq x_i\}|$ for all $a \in \mathcal{A}$, and repeating, for $i = 1, \ldots, n$, the following two steps: (i) $x_i^* \leftarrow$ a maximal $y \in \mathcal{P}_i \cap x_i^+$ such that $y \not\succeq a_i$ for all $a \in \mathcal{A}$ with $c(a) = 1$ and $a_i \not\preceq x_i$; (ii) $c(a) \leftarrow c(a) - 1$ for each $a \in \mathcal{A}$ such that $a_i \preceq x_i^*$.

It is also worth noting that, if condition (4) does not hold, then the problem becomes NP-hard in general, even if $\mathcal{P} = \{0, 1\}^n$ is just the Boolean cube [EG95].

Given any $\mathcal{Q} \subseteq \mathcal{P}$, let us denote by

$$\mathcal{A}(\mathcal{Q}) = \{a \in \mathcal{A} \mid a^+ \cap \mathcal{Q} \neq \emptyset\}, \qquad \mathcal{B}(\mathcal{Q}) = \{b \in \mathcal{B} \mid b^- \cap \mathcal{Q} \neq \emptyset\}.$$

Note that, for $a \in \mathcal{A}$ and $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_n$, $a^+ \cap \mathcal{Q} \neq \emptyset$ if and only if $a_i^+ \cap \mathcal{Q}_i \neq \emptyset$, for all $i \in [n]$. Thus, the sets $\mathcal{A}(\mathcal{Q})$ and $\mathcal{B}(\mathcal{Q})$ can be found in $O(nm\mu)$ time.[1] A simple but important observation, which will be used frequently in the algorithms below, is that

$$\mathcal{Q} \subseteq \mathcal{A}^+ \cup \mathcal{B}^- \iff \mathcal{Q} \subseteq \mathcal{A}(\mathcal{Q})^+ \cup \mathcal{B}(\mathcal{Q})^-. \tag{5}$$

To solve problem DUAL$(\mathcal{P}, \mathcal{A}, \mathcal{B})$, we shall use the same general approach used in [FK96] to solve the hypergraph dualization problem, by decomposing it into a number of smaller subproblems which are solved recursively. In each such subproblem, we start with a subposet $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_n \subseteq \mathcal{P}$ (initially $\mathcal{Q} = \mathcal{P}$) and two subsets $\mathcal{A}(\mathcal{Q}) \subseteq \mathcal{A}$ and $\mathcal{B}(\mathcal{Q}) \subseteq \mathcal{B}$, and we want to check whether $\mathcal{A}(\mathcal{Q})$ and $\mathcal{B}(\mathcal{Q})$ are dual in $\mathcal{Q}$, i.e., whether $\mathcal{Q} \subseteq \mathcal{A}(\mathcal{Q})^+ \cup \mathcal{B}(\mathcal{Q})^-$. Note that since $\mathcal{B} \subseteq \mathcal{I}(\mathcal{A})$ is assumed, (4) continues to hold for the recursive subproblems. The decomposition of $\mathcal{Q}$ is done by decomposing one factor poset, say $\mathcal{Q}_i$, into a number of (not necessarily disjoint) subposets $\mathcal{Q}_i^1, \ldots, \mathcal{Q}_i^r$ and solving r subproblems on the r different posets $\mathcal{Q}_1 \times \cdots \times \mathcal{Q}_{i-1} \times \mathcal{Q}_i^j \times \mathcal{Q}_{i+1} \times \cdots \times \mathcal{Q}_n$, $j = 1, \ldots, r$. In most of the cases, a number of decomposition rules may be followed, based on the sizes of certain subsets of $\mathcal{A}$ and $\mathcal{B}$, with the objective of reducing the problem size from one level of the recursion to the next. To estimate this reduction in size (only in the analysis of the running time), we measure the change in the "*volume*" of the problem defined as $v = v(\mathcal{A}, \mathcal{B}) \stackrel{\text{def}}{=} |\mathcal{A}||\mathcal{B}|$ (actually, in section 6, we use $v(\mathcal{Q}, \mathcal{A}, \mathcal{B}) = |\mathcal{A}||\mathcal{B}| \sum_{i=1}^n |\mathcal{Q}_i|$). For brevity, we shall denote by $\overline{\mathcal{Q}}^i$ the product $\mathcal{Q}_1 \times \cdots \times \mathcal{Q}_{i-1} \times \mathcal{Q}_{i+1} \times \cdots \times \mathcal{Q}_n$, and accordingly by $\overline{q}^i$ the vector $(q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n)$, for an element $q = (q_1, q_2, \ldots, q_n) \in \mathcal{Q}$. When the index i is understood from the context, we will use $\overline{\mathcal{Q}}$ and $\bar{q}$ for simplicity.

[1] In fact, by the way $\mathcal{Q}$ is chosen in our algorithms, these sets can be found in $O(nm)$ time.

Procedure PD($\mathcal{Q}, \mathcal{A}, \mathcal{B}$)**:**
Input: A subposet $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_n \subseteq \mathcal{P}$ and two antichains $\mathcal{A}, \mathcal{B} \subseteq \mathcal{P}$
Output: **true** if $\mathcal{Q} \subseteq (\mathcal{A}^+ \cup \mathcal{B}^-)$ and **false** otherwise

1. $\mathcal{A} \leftarrow \mathcal{A}(\mathcal{Q})$, $\mathcal{B} \leftarrow \mathcal{B}(\mathcal{Q})$
2. $\mathcal{A} \leftarrow$ PROJECT($\mathcal{Q}, \mathcal{A}$), $\mathcal{B} \leftarrow$ PROJECT($\mathcal{Q}, \mathcal{B}$)
3. **if** $\min\{|\mathcal{A}|, |\mathcal{B}|\} \leq Const.$ **then**
4. **return** POLY-DUAL($\mathcal{Q}, \mathcal{A}, \mathcal{B}$)
5. Using the appropriate decomposition rule, select $i \in [n]$,
and decompose $\mathcal{Q}_i$ into $\mathcal{Q}_i^1, \ldots, \mathcal{Q}_i^r$
6. **return** $\bigwedge_{j=1}^{r}$ PD($\mathcal{Q}_i^j \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}$)

FIG. 4. *The general dualization procedure.*

A general high-level dualization procedure is shown in Figure 4. In this procedure, we use 2 subroutines: PROJECT and POLY-DUAL. The second of these routines acts as the base case for recursion, while the first is used to ensure that, at that base level, the subsets $\mathcal{A}, \mathcal{B}, \mathcal{Q} \subseteq \mathcal{P}$ satisfy $\mathcal{A}, \mathcal{B} \subseteq \mathcal{Q}$. The reason that we need the latter condition and the details of these two routines will be given in subsections 3.3 and 3.4, respectively. In the next section, we derive some general decomposition rules that can be used in step 5 of the procedure. The selection of which decomposition rule to use in the algorithm depends on the *frequencies* of the element, at which the decomposition is performed, with respect to $\mathcal{A}$ and $\mathcal{B}$, but does not otherwise assume anything about these frequencies. Assuming duality of $\mathcal{A}$ and $\mathcal{B}$, one can show that there exists a "high-frequency" element in one of the factor posets. Using this element for decomposition at each recursion level usually yields much simpler algorithms, but with worse running times with respect to m, although possibly better in terms of the other parameters (e.g., width). In fact, this is the only method we know of for getting quasi-polynomial bounds in the width, in the case of products lattices of intervals (see section 6). In subsection 3.2.4, we give the arguments for the existence of such high-frequency elements.

We assume that procedure PD($\mathcal{Q}, \mathcal{A}, \mathcal{B}$) returns either true or false depending on whether $\mathcal{A}$ and $\mathcal{B}$ are dual in $\mathcal{Q}$ or not. Returning a vector $x \in \mathcal{Q} \setminus (\mathcal{A}^+ \cup \mathcal{B}^-)$ in the latter case is straightforward, as it can be obtained from any subproblem that failed the test for duality.

In the rest of this paper, we shall denote by $C(v(\mathcal{Q}, \mathcal{A}, \mathcal{B}))$ the total number of recursive subproblems created by procedure PD($\mathcal{Q}, \mathcal{A}, \mathcal{B}$) on a problem of size $v(\mathcal{Q}, \mathcal{A}, \mathcal{B})$. We assume that $C(v) \leq R(v)$, where $R(v)$ is a *superadditive*[2] function of v (i.e., $R(v) + R(v') \leq R(v + v')$ for all $v, v' \geq 0$), is monotone (i.e., $f(v) \geq f(v')$ for all $v \geq v'$), $R(0) = 0$, and $R(1) \geq 1$. Thus we may assume also, without loss of generality, that $C(v)$ is monotone and superadditive.

3.2. Decomposition.

3.2.1. Independent decomposition. Let us call two subposets $\mathcal{Q}, \mathcal{R} \subseteq \mathcal{P}$ *independent* if $q \not\leq r$ and $q \not\geq r$ for all $q \in \mathcal{Q}, r \in \mathcal{R}$. The following decomposition rule can be used to reduce the problem on products of forests into one in which each forest has exactly on connected component, i.e., a tree.

[2]This is naturally satisfied by any monotone superlinear function.

PROPOSITION 2. *Let $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_n$ and $\mathcal{A}, \mathcal{B} \subseteq \mathcal{Q}$. Suppose that poset $\mathcal{Q}_i$ can be partitioned into two independent posets $\mathcal{Q}_i'$ and $\mathcal{Q}_i''$. Let $\mathcal{Q}' = \mathcal{Q}_i' \times \overline{\mathcal{Q}}$, $\mathcal{Q}'' = \mathcal{Q}_i'' \times \overline{\mathcal{Q}}$, $\mathcal{A}' = \mathcal{A}(\mathcal{Q}')$, $\mathcal{B}' = \mathcal{B}(\mathcal{Q}')$, $\mathcal{A}'' = \mathcal{A}(\mathcal{Q}'')$, and $\mathcal{B}'' = \mathcal{B}(\mathcal{Q}'')$. If $C(v(\mathcal{A}', \mathcal{B}')) \leq R(v(\mathcal{A}', \mathcal{B}'))$ and $C(v(\mathcal{A}'', \mathcal{B}'')) \leq R(v(\mathcal{A}'', \mathcal{B}''))$, then $C(v(\mathcal{A}, \mathcal{B})) \leq R(v(\mathcal{A}, \mathcal{B}))$.*

Proof. We observe by (5) and the independence of $\mathcal{Q}', \mathcal{Q}''$ that

$$\mathcal{Q} \subseteq \mathcal{A}^+ \cup \mathcal{B}^- \iff \mathcal{Q}' \subseteq (\mathcal{A}')^+ \cup (\mathcal{B}')^- \text{ and } \mathcal{Q}'' \subseteq (\mathcal{A}'')^+ \cup (\mathcal{B}'')^-.$$

Clearly, if $\mathcal{A}' \cup \mathcal{B}' = \emptyset$ (or $\mathcal{A}'' \cup \mathcal{B}'' = \emptyset$), then any element in $\mathcal{Q}'$ (respectively, in $\mathcal{Q}''$) does not belong to $\mathcal{A}^+ \cup \mathcal{B}^-$. On the other hand, if these unions are not empty, then we can proceed by recursively solving the two subproblems DUAL$(\mathcal{Q}', \mathcal{A}', \mathcal{B}')$ and DUAL$(\mathcal{Q}'', \mathcal{A}'', \mathcal{B}'')$. This gives

$$C(v(\mathcal{A}, \mathcal{B})) = 1 + C(v(\mathcal{A}', \mathcal{B}')) + C(v(\mathcal{A}'', \mathcal{B}'')) \leq 1 + R(v(\mathcal{A}', \mathcal{B}')) + R(v(\mathcal{A}'', \mathcal{B}'')).$$

Note that $\{\mathcal{A}', \mathcal{A}''\}$ and $\{\mathcal{B}', \mathcal{B}''\}$ form partitions of $\mathcal{A}$ and $\mathcal{B}$, respectively, and therefore, we get by the superadditivity and monotonicity of $R(\cdot)$

$$\begin{aligned} R(v(\mathcal{A}, \mathcal{B})) &= R(v(\mathcal{A}', \mathcal{B}') + v(\mathcal{A}'', \mathcal{B}'') + v(\mathcal{A}', \mathcal{B}'') + v(\mathcal{A}'', \mathcal{B}')) \\ &\geq R(v(\mathcal{A}', \mathcal{B}')) + R(v(\mathcal{A}'', \mathcal{B}'')) + R(v(\mathcal{A}', \mathcal{B}'')) + R(v(\mathcal{A}'', \mathcal{B}')) \\ &\geq R(v(\mathcal{A}', \mathcal{B}')) + R(v(\mathcal{A}'', \mathcal{B}'')) + 1, \end{aligned}$$

where the last inequality follows from the fact that if $v(\mathcal{A}, \mathcal{B}) > 0$, then either $v(\mathcal{A}', \mathcal{B}'') > 0$ or $v(\mathcal{A}'', \mathcal{B}') > 0$. ∎

3.2.2. General decomposition. For an operator $\circ \in \{\preceq, \not\preceq, \succeq, \not\succeq\}$, a subset $\mathcal{X} \subseteq \mathcal{P}$, an $i \in [n]$, and an element $z \in \mathcal{P}_i$, we use the notation $\mathcal{X}_\circ(z) \stackrel{\text{def}}{=} \{x \in \mathcal{X} : x_i \circ z\}$. In general, the algorithms will decompose a given problem by selecting an $i \in [n]$ and partitioning $\mathcal{Q}_i$ into two subposets $\mathcal{Q}_i'$ and $\mathcal{Q}_i''$, defining accordingly two poset products $\mathcal{Q}'$ and $\mathcal{Q}''$. Specifically, let $a^o \in \mathcal{A}$, $b^o \in \mathcal{B}$ be arbitrary elements of $\mathcal{A}, \mathcal{B}$ (in fact the algorithm in section 4.1 will select specific elements $a^o \in \mathcal{A}$ and $b^o \in \mathcal{B}$). By (4), there exists an $i \in [n]$ such that $a_i^o \not\preceq b_i^o$. Let $\mathcal{Q}_i' \leftarrow \mathcal{Q}_i \cap (a_i^o)^+$, $\mathcal{Q}_i'' \leftarrow \mathcal{Q}_i \setminus \mathcal{Q}_i'$ (we may alternatively set $\mathcal{Q}_i'' \leftarrow \mathcal{Q}_i \cap (b_i^o)^-$ and $\mathcal{Q}_i' \leftarrow \mathcal{Q}_i \setminus \mathcal{Q}_i''$; see section 5.1). Defining $\mathcal{Q}' = \mathcal{Q}_i' \times \overline{\mathcal{Q}}$ and $\mathcal{Q}'' = \mathcal{Q}_i'' \times \overline{\mathcal{Q}}$ to be the two subposets induced by this partitioning and letting $\mathcal{A}' \stackrel{\text{def}}{=} \mathcal{A}(\mathcal{Q}') = \mathcal{A}_\succeq(a_i^o)$, $\mathcal{A}'' \stackrel{\text{def}}{=} \mathcal{A}(\mathcal{Q}'') = \mathcal{A}_{\not\succeq}(a_i^o)$, $\mathcal{B}' \stackrel{\text{def}}{=} \mathcal{B}(\mathcal{Q}') = \mathcal{B}_\succeq(a_i^o)$, $\mathcal{B}'' \stackrel{\text{def}}{=} \mathcal{B}(\mathcal{Q}'') = \mathcal{B}_{\not\succeq}(a_i^o)$, we conclude by (5) that $\mathcal{Q} \subseteq \mathcal{A}^+ \cup \mathcal{B}^-$ *if and only if*

$$\mathcal{Q}' \subseteq \mathcal{A}^+ \cup (\mathcal{B}')^- \quad \text{and} \tag{6}$$

$$\mathcal{Q}'' \subseteq (\mathcal{A}'')^+ \cup \mathcal{B}^-. \tag{7}$$

Thus we have decomposed the original problem into two new subproblems. Note that the volumes of the resulting problems are strictly less than the volume of the original problem. For lattices and forests, it may be necessary to further decompose the subposet $\mathcal{Q}_i''$ in order to maintain a certain nice property (lattice property, connectedness of the precedence graph) which allows for the *projection step* described in section 3.4.

Clearly, there may exist precedence relations between the elements of $\mathcal{Q}_i'$ and $\mathcal{Q}_i''$ and, therefore, the two subproblems (6) and (7) may not be independent. Once we

get an affirmative answer to one subproblem, we gain some information about the solution of the other. The following lemma shows how to utilize this dependence to further decompose the other subproblem in such a case.

LEMMA 1. *Given $z \in \mathcal{Q}_i$, let $\mathcal{R}'_i = \mathcal{Q}_i \cap z^+$, $\mathcal{R}''_i \subseteq \mathcal{Q}_i \cap z^- \setminus \{z\}$ be two disjoint subposets of $\mathcal{Q}_i$. Define*

$$\mathcal{A}^2 = \{a \in \mathcal{A} \mid a_i^+ \cap \mathcal{R}''_i \neq \emptyset\}, \qquad \mathcal{A}^1 = \{a \in \mathcal{A} \setminus \mathcal{A}^2 |\ a_i^+ \cap \mathcal{R}'_i \neq \emptyset\},$$
$$\mathcal{B}^1 = \{b \in \mathcal{B} \mid b_i^- \cap \mathcal{R}'_i \neq \emptyset\}, \qquad \mathcal{B}^2 = \{b \in \mathcal{B} \setminus \mathcal{B}^1 |\ b_i^- \cap \mathcal{R}''_i \neq \emptyset\}.$$

Suppose further that $\mathcal{R}'_i \times \overline{\mathcal{Q}} \subseteq (\mathcal{A}^2 \cup \mathcal{A}^1)^+ \cup (\mathcal{B}^1)^-$; then

$$\mathcal{R}''_i \times \overline{\mathcal{Q}} \subseteq (\mathcal{A}^2)^+ \cup (\mathcal{B}^1 \cup \mathcal{B}^2)^- \iff \forall a \in \mathcal{A}_{\preceq}(z) : \mathcal{R}''_i \times (\overline{\mathcal{Q}} \cap \overline{a}^+) \subseteq (\mathcal{A}^2)^+ \cup (\mathcal{B}^2)^-.$$

Proof. Suppose first that $\mathcal{R}''_i \times \overline{\mathcal{Q}} \subseteq (\mathcal{A}^2)^+ \cup (\mathcal{B}^1 \cup \mathcal{B}^2)^-$. Let $(q_i, \overline{q}) \in \mathcal{R}''_i \times (\overline{\mathcal{Q}} \cap \overline{a}^+)$ for some $a \in \mathcal{A}_{\preceq}(z)$; then $(q_i, \overline{q}) \in (\mathcal{A}^2)^+ \cup (\mathcal{B}^1 \cup \mathcal{B}^2)^-$. If $(q_i, \overline{q}) \preceq (b_1, \overline{b}) \in \mathcal{B}^1$, then by the definition of $\mathcal{B}^1$, there is a $y \in \mathcal{R}'_i$ such that $y \preceq b_i$. But then, $a \in \mathcal{A}_{\preceq}(z)$, $\overline{q} \in \overline{\mathcal{Q}} \cap \overline{a}^+$, and $y \in \mathcal{R}'_i$ imply that $(a_i, \overline{a}) \preceq (z, \overline{q}) \preceq (y, \overline{q}) \preceq (b_i, \overline{b})$, which contradicts the assumed condition (4). This shows that $(q_i, \overline{q}) \in (\mathcal{A}^2)^+ \cup (\mathcal{B}^2)^-$.

For the other direction, let $(q_i, \overline{q}) \in (\mathcal{R}''_i \times \overline{\mathcal{Q}}) \setminus (\mathcal{B}^1)^-$. Since $x \preceq y$ for all $x \in \mathcal{R}''_i$, $y \in \mathcal{R}'_i$, we must have $(y, \overline{q}) \notin (\mathcal{B}^1)^-$ for all $y \in \mathcal{R}'_i$, for otherwise we get the contradiction $(q_1, \overline{q}) \preceq (y, \overline{q}) \preceq (b_1, \overline{b})$ for some $b \in \mathcal{B}^1$. Now we use our assumption that $\mathcal{R}'_i \times \overline{\mathcal{Q}} \subseteq (\mathcal{A}^1 \cup \mathcal{A}^2)^+ \cup (\mathcal{B}^1)^-$ to conclude that $(y, \overline{q}) \in (\mathcal{A}^1 \cup \mathcal{A}^2)^+$ for all $y \in \mathcal{R}'_i$. In particular, we have $(z, \overline{q}) \succeq (a_1, \overline{a})$ for some $(a_1, \overline{a}) \in \mathcal{A}^1 \cup \mathcal{A}^2$. But this implies that $a \in \mathcal{A}_{\preceq}(z)$ and hence that $(q_1, \overline{q}) \in \mathcal{R}''_i \times (\overline{\mathcal{Q}} \cap \overline{a}^+)$ for some $a \in \mathcal{A}_{\preceq}(z)$. This gives $(q_i, \overline{q}) \in (\mathcal{A}^2)^+ \cup (\mathcal{B}^2)^-$. ∎

By considering the dual poset of $\mathcal{P}$ (that is, the poset $\mathcal{P}^*$ with the same set of elements as $\mathcal{P}$, but such that $x \prec y$ in $\mathcal{P}^*$ whenever $x \succ y$ in $\mathcal{P}$) and exchanging the roles of $\mathcal{A}$ and $\mathcal{B}$, we get the following symmetric version of Lemma 1.

LEMMA 2. *Let $\mathcal{R}''_i = \mathcal{Q}_i \cap z^-$, $\mathcal{R}'_i \subseteq \mathcal{Q}_i \cap z^+ \setminus \{z\}$ be two disjoint subposets of $\mathcal{Q}_i$ where $z \in \mathcal{Q}_i$. Let $\mathcal{A}^1, \mathcal{A}^2, \mathcal{B}^1, \mathcal{B}^2$ be defined as in Lemma 1. Suppose that $\mathcal{R}''_i \times \overline{\mathcal{Q}} \subseteq (\mathcal{A}^2)^+ \cup (\mathcal{B}^1 \cup \mathcal{B}^2)^-$; then*

$$\mathcal{R}'_i \times \overline{\mathcal{Q}} \subseteq (\mathcal{A}^1 \cup \mathcal{A}^2)^+ \cup (\mathcal{B}^1)^- \iff \forall b \in \mathcal{B}_{\succeq}(z) : \mathcal{R}'_i \times (\overline{\mathcal{Q}} \cap \overline{b}^-) \subseteq (\mathcal{A}^1)^+ \cup (\mathcal{B}^1)^-.$$

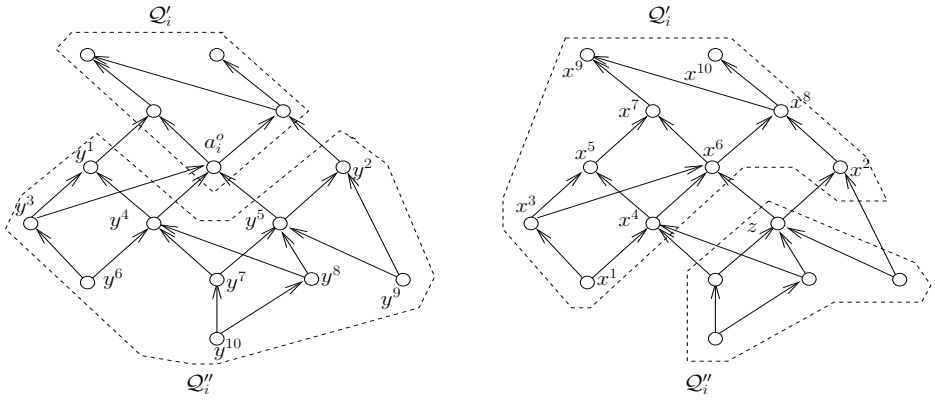
We now use Lemma 1 inductively to get a further decomposition of poset $\mathcal{Q}''_i$. Suppose that one of the subproblems, say (6), has no solution, i.e., $\mathcal{Q}' \subseteq \mathcal{A}^+ \cup (\mathcal{B}')^-$. Then we can proceed in this case as follows. Let us use $y^1, \ldots, y^k$ to denote the elements of $\mathcal{Q}''_i$ and assume, without loss of generality, that they are *inversely topologically sorted* in this order, that is, $y^j \prec y^r$ implies $j > r$ (see Figure 5(a)). Let us decompose (7) further into the k subproblems

$$\{y^j\} \times \overline{\mathcal{Q}} \subseteq (\mathcal{A}''_{\preceq}(y^j))^+ \cup \left[\mathcal{B}''_{=}(y^j) \cup \left(\bigcup_{x \in (y^j)^\top} \mathcal{B}_{\succeq}(x)\right)\right]^-, \quad j = 1, \ldots, k. \tag{8}$$

The following lemma will allow us to eliminate the contribution of the set $\mathcal{B}'$ in subproblems (8) at the expense of possibly introducing at most $|\mathcal{A}|^{W(\mathcal{Q})}$ additional subproblems.

LEMMA 3. *Given $y^j \in \mathcal{Q}''_i$, suppose we know that $(x^+ \cap \mathcal{Q}_i) \times \overline{\mathcal{Q}} \subseteq \mathcal{A}^+ \cup (\mathcal{B}^{\succeq}(x))^-$ for all $x \in (y^j)^\top$. Then (8) is equivalent to*

$$\{y^j\} \times \left[\overline{\mathcal{Q}} \cap \left(\bigcap_{x \in (y^j)^\top} \overline{a}(x)^+\right)\right] \subseteq (\mathcal{A}''_{\preceq}(y^j))^+ \cup (\mathcal{B}''_{=}(y^j))^- \tag{9}$$



(a) Rules $(R1)$ and $(R2)$. (b) Rule $(R2')$.

FIG. 5. *Decomposing the poset* $\mathcal{Q}_i$.

for all collections $\{a(x) \in \mathcal{A}_{\preceq}(x) \mid x \in (y^j)^\top\}$ *for* $j = 1, \ldots, k$. *(That is, if* $(y^j)^\top = \{x^1, \ldots, x^s\}$, *then we consider all collections of the form* $\{a(x^1), \ldots, a(x^s)\}$, *where* $a(x^1) \in \mathcal{A}_{\preceq}(x^1), \ldots, a(x^s) \in \mathcal{A}_{\preceq}(x^s)$.*)*

Proof. We prove by induction on $|X|$, where $X \subseteq (y^j)^\top$, that

(10)

$$\{y^j\} \times \overline{\mathcal{Q}} \subseteq (\mathcal{A}''_{\preceq}(y^j))^+ \cup \left[\mathcal{B}''_{=}(y^j) \cup \left(\bigcup_{x \in (y^j)^\top} \mathcal{B}_{\succeq}(x)\right)\right]^-$$

$$\iff \{y^j\} \times \left[\overline{\mathcal{Q}} \cap \left(\bigcap_{x \in X} \overline{a}(x)^+\right)\right] \subseteq (\mathcal{A}''_{\preceq}(y^j))^+ \cup \left[\mathcal{B}''_{=}(y^j) \cup \left(\bigcup_{x \in (y^j)^\top \setminus X} \mathcal{B}_{\succeq}(y)\right)\right]^-$$

for all collections $\{a(y) \in \mathcal{A}_{\preceq}(x) \mid x \in X\}$. This trivially holds for $X = \emptyset$ and will prove the lemma for $X = (y^j)^\top$. To show (10), assume that it holds for some $X \subset (y^j)^\top$, and let $u \in (y^j)^\top \setminus Y$. Consider a subproblem of the form

$$\{y^j\} \times \left[\overline{\mathcal{Q}} \cap \left(\bigcap_{x \in X} \overline{a}(x)^+\right)\right]$$

$$\subseteq (\mathcal{A}''_{\preceq}(y^j))^+ \cup \left[\mathcal{B}''_{=}(y^j) \cup \mathcal{B}_{\succeq}(u) \cup \left(\bigcup_{x \in (y^j)^\top \setminus (X \cup \{u\})} \mathcal{B}_{\succeq}(y)\right)\right]^-$$

for some collection $\{a(y) \in \mathcal{A}_{\preceq}(x) \mid x \in X\}$. Now we apply Lemma 1 with $z \leftarrow u$, $\mathcal{R}'_i \leftarrow z^+ \cap \mathcal{Q}_i$, $\mathcal{R}''_i \leftarrow \{y^j\}$, $\mathcal{A}^2 \leftarrow \mathcal{A}_{\preceq}(y^j)$, $\mathcal{B}^1 \leftarrow \mathcal{B}_{\succeq}(u)$, and $\mathcal{B}^2 \leftarrow \mathcal{B}''_{=}(y^j) \cup (\bigcup_{x \in (y^j)^\top \setminus (X \cup \{u\})} \mathcal{B}_{\succeq}(y))$ to get the required result. ∎

Informally, Lemma 3 says that, given $y^j \in \mathcal{Q}''_i$, if the dualization subproblems for all subposets that lie above y^j have been already verified to have no solution, then we can solve subproblem (8) by solving at most $\prod_{x \in (y^j)^\top} |\mathcal{A}_{\preceq}(x)|$ subproblems of the form (9). Observe that it is important to check subproblems (8) in the reverse topological order $j = 1, \ldots, h$ in order to be able to use Lemma 3.

3.2.3. Decomposition rules. Using the decomposition lemmas stated in the previous subsection, we now derive some general decomposition rules that will be used later by the algorithms.

Rule (R1). Solve the two subproblems (corresponding to) (6) and (7).

Rule (R2). Solve subproblem (6). If it has a solution, then we get an element $x \in \mathcal{Q} \setminus (\mathcal{A}^+ \cup \mathcal{B}^-)$. Otherwise, we solve subproblems (9) for all collections $\{a(x) \in \mathcal{A}_{\preceq}(x) \mid x \in (y^j)^\top\}$, for $j = 1, \ldots, k$, where $y^1, \ldots, y^h$ denote the elements of $\mathcal{Q}_i''$ in reverse topological order (see Figure 5(a)).

Suppose, finally, that we decompose $\mathcal{Q}_i$ by selecting an element $z \in \mathcal{Q}_i$, letting $\mathcal{Q}_i'' \leftarrow \mathcal{Q}_i \cap z^-$, $\mathcal{Q}_i' \leftarrow \mathcal{Q}_i \setminus z^-$, $\mathcal{A}'' = \mathcal{A}_{\preceq}(z)$, $\mathcal{A}' = \mathcal{A} \setminus \mathcal{A}''$, and $\mathcal{B}' = \mathcal{B}_{\succeq}(z)$. By exchanging the roles of $\mathcal{A}$ and $\mathcal{B}$ and replacing $\mathcal{P}$ by its dual poset $\mathcal{P}^*$ in Rule ($R2$) above, we can also arrive at the following symmetric version of this rule (see Figure 5(b)).

Rule (R2′). Solve subproblem (7). If it does not have a solution, then solve the subproblems

$$\{x^j\} \times \left[\overline{\mathcal{Q}} \cap \left(\bigcap_{y \in (x^j)^\perp} \overline{b}(y)^- \right) \right] \subseteq (\mathcal{A}'_{=}(x^j))^+ \cup (\mathcal{B}'_{\succeq}(x^j))^- \tag{11}$$

for all collections $\{b(y) \in \mathcal{B}_{\succeq}(y) \mid y \in (x^j)^\perp\}$, for $j = 1, \ldots, k$, where $x^1, \ldots, x^k$ denote the elements of $\mathcal{Q}_i'$ in *topological* order (that is, $x^j \prec x^r$ implies $j < r$).

In sections 4, 5, and 6, we show how to use the above rules for decomposing a given dualization problem into smaller subproblems. The algorithms will select between these rules in such a way that the total volume is reduced significantly from one recursion level to the next.

3.2.4. High-frequency based decomposition. Assume that $\mathcal{A}, \mathcal{B}$ satisfy (4), and let us denote by $\mathrm{Min}(\mathcal{Q}_i)$ and $\mathrm{Max}(\mathcal{Q}_i)$, respectively, the sets of minimal and maximal elements of poset $\mathcal{Q}_i$. Define the *support* of $a \in \mathcal{A}$ (respectively, $b \in \mathcal{B}$) to be the set of all nonminimum coordinates of a (respectively, the set of all nonmaximum coordinates of b):

$$\mathrm{Supp}(a) = \{i \in [n] \ : \ \mathrm{Min}(\mathcal{Q}_i) \neq \{a_i\}\}, \qquad \mathrm{Supp}(b) = \{i \in [n] \ : \ \mathrm{Max}(\mathcal{Q}_i) \neq \{b_i\}\}.$$

Let $\alpha = \alpha(\mathcal{Q}) \stackrel{\mathrm{def}}{=} \max_{i \in [n]} \{|\mathrm{Min}(\mathcal{Q}_i) \cup \mathrm{Max}(\mathcal{Q}_i)|\}$. The following lemma generalizes a known fact for dual Boolean functions (cf. [FK96]).

LEMMA 4. *If $\mathcal{A}, \mathcal{B}$ are dual in $\mathcal{Q}$, then there exists an element $x \in \mathcal{A} \cup \mathcal{B}$ with a logarithmically small support: $|\mathrm{Supp}(x)| \leq \alpha \ln m$, where $m = |\mathcal{A}| + |\mathcal{B}|$.*

Proof. Let $z \in \mathcal{Q}$ be the vector obtained by picking each coordinate z_i randomly from $\mathcal{X}_i \stackrel{\mathrm{def}}{=} \mathrm{Min}(\mathcal{Q}_i) \cup \mathrm{Max}(\mathcal{Q}_i)$, $i = 1, \ldots, n$, and consider the random variable $N(z) \stackrel{\mathrm{def}}{=} |\{a \in \mathcal{A} \mid z \succeq a\}| + |\{b \in \mathcal{B} \mid z \preceq b\}|$. Then the expected value of $N(z)$ is given by

$$\begin{aligned}
\mathbb{E}[N(z)] &= \sum_{a \in \mathcal{A}} \Pr\{z \succeq a\} + \sum_{b \in \mathcal{B}} \Pr\{z \preceq b\} \\
&= \sum_{a \in \mathcal{A}} \prod_{i \in \mathrm{Supp}(a)} \tfrac{|\mathcal{X}_i \cap a_i^+|}{|\mathcal{X}_i|} + \sum_{b \in \mathcal{B}} \prod_{i \in \mathrm{Supp}(b)} \tfrac{|\mathcal{X}_i \cap b_i^-|}{|\mathcal{X}_i|} \\
&\leq \sum_{a \in \mathcal{A}} \prod_{i \in \mathrm{Supp}(a)} \left(1 - \tfrac{1}{|\mathcal{X}_i|}\right) + \sum_{b \in \mathcal{B}} \prod_{i \in \mathrm{Supp}(b)} \left(1 - \tfrac{1}{|\mathcal{X}_i|}\right) \\
&\leq \sum_{a \in \mathcal{A}} \left(1 - \tfrac{1}{\alpha}\right)^{|\mathrm{Supp}(a)|} + \sum_{b \in \mathcal{B}} \left(1 - \tfrac{1}{\alpha}\right)^{|\mathrm{Supp}(b)|}.
\end{aligned} \tag{12}$$

Clearly $\mathbb{E}[N(z)] \geq 1$, for otherwise there exists an element $x \in \mathcal{L} \setminus (\mathcal{A}^+ \cup \mathcal{B}^-)$ (which can be found in $O(nm\alpha)$ using the standard method of conditional expectation [MR95]). Let $r = \min\{|\operatorname{Supp}(z)| \ : \ z \in \mathcal{A} \cup \mathcal{B}\}$. Then (12) implies that

$$1 \leq \mathbb{E}[N(z)] \leq (|\mathcal{A}| + |\mathcal{B}|)\left(1 - \frac{1}{\alpha}\right)^r \leq me^{-r/\alpha}.$$

The lemma follows. ∎

Next we show that, for any dual pair $(\mathcal{A}, \mathcal{B})$, a high-frequency element exists with respect to either $\mathcal{A}$ or $\mathcal{B}$.

LEMMA 5. *Let $\mathcal{A}, \mathcal{B}$ be a pair of dual subsets of $\mathcal{Q}$ with $|\mathcal{A}||\mathcal{B}| \geq 1$. Then there exist a coordinate $i \in [n]$ and an element $z \in \mathcal{Q}_i$, such that either*

(i) $|\mathcal{A}_{\succeq}(z)| \geq 1$ *and* $|\mathcal{B}_{\not\preceq}(z)| \geq \frac{|\mathcal{B}|}{\alpha(\mathcal{Q}) \ln m}$, *or*

(ii) $|\mathcal{B}_{\preceq}(z)| \geq 1$ *and* $|\mathcal{A}_{\not\succeq}(z)| \geq \frac{|\mathcal{A}|}{\alpha(\mathcal{Q}) \ln m}$.

There also exist a coordinate $i \in [n]$ and an element $z \in \mathcal{Q}_i$, such that either

(iii) $|\mathcal{A}_{\not\succeq}(z)| \geq 1$ *and* $|\mathcal{B}_{\preceq}(z)| \geq \frac{|\mathcal{B}|}{\alpha(\mathcal{Q}) W(\mathcal{Q}_i) \ln m}$, *or*

(iv) $|\mathcal{B}_{\not\preceq}(z)| \geq 1$ *and* $|\mathcal{A}_{\succeq}(z)| \geq \frac{|\mathcal{A}|}{\alpha(\mathcal{Q}) W(\mathcal{Q}_i) \ln m}$.

Proof. By Lemma 4, $\mathcal{A} \cup \mathcal{B}$ contains an element x with $|\operatorname{Supp}(x)| \leq \alpha \ln m$. Suppose, without loss of generality, that $x \in \mathcal{A}$. From condition (4), we know that for every $b \in \mathcal{B}$, there is an $i \in \operatorname{Supp}(b) \cap \operatorname{Supp}(x)$ such that $b_i \not\succeq x_i$. Thus

$$|\mathcal{B}| = \left| \bigcup_{i \in \operatorname{Supp}(x)} \mathcal{B}_{\not\succeq}(x_i) \right| \leq \sum_{i \in \operatorname{Supp}(x)} |\mathcal{B}_{\not\succeq}(x_i)|,$$

and therefore there is an $i \in [n]$ such that $|\mathcal{B}_{\not\succeq}(x_i)| \geq |\mathcal{B}|/|\operatorname{Supp}(x)| \geq |\mathcal{B}|/(\alpha \ln m)$, which implies (i) for $z = x_i$.

To show (iii), consider the set $\mathcal{Y} = \mathcal{I}(\{x_i\})$ of maximal independent elements in $\mathcal{Q}_i \setminus \{x_i\}^+$, and observe that

$$|\mathcal{B}_{\not\succeq}(x_i)| = \left| \bigcup_{z \in \mathcal{Y}} \mathcal{B}_{\preceq}(z) \right| \leq \sum_{z \in \mathcal{Y}} |\mathcal{B}_{\preceq}(z)|.$$

Noting that $|\mathcal{Y}| \leq W(\mathcal{Q}_i)$, we conclude that (iii) holds. If x actually belongs to $\mathcal{B}$, then by a similar argument we obtain (ii) and (iv). ∎

3.3. Polynomial dualization when one of the sets is small. When one of the sets $\mathcal{A}$ or $\mathcal{B}$ has constant size, the problem can be solved in polynomial time.

PROPOSITION 3. *Suppose that $\min\{|\mathcal{A}|, |\mathcal{B}|\} \leq k$, $\mathcal{A}, \mathcal{B} \subseteq \mathcal{P}$; then problem DUAL($\mathcal{P}$, $\mathcal{A}, \mathcal{B}$) is solvable in time $O(n^{k+1}\ mW(\mathcal{P})^{k+1}\mu(\mathcal{P}))$.*

Proof. Let us assume, without loss of generality, that $\mathcal{B} = \{b^1, \ldots, b^k\}$ for some constant k. Then problem DUAL($\mathcal{P}$, $\mathcal{A}, \mathcal{B}$) can be reduced to n^k subproblems of the form DUAL($\mathcal{P}', \mathcal{A}, \emptyset$), where $\mathcal{P}' = \mathcal{P}'_1 \times \cdots \times \mathcal{P}'_n$, is obtained from $\mathcal{P}$ by selecting, for each $j \in [k]$, a coordinate $i_j \in [n]$ and setting $\mathcal{P}'_i = \mathcal{P}_i \setminus \bigcup_{j \in [k]} (b^j_{i_j})^-$.

Clearly, $\mathcal{P}' \subseteq \mathcal{A}^+$ if and only if $\operatorname{Min}(\mathcal{P}') \subseteq \mathcal{A}^+$, where $\operatorname{Min}(\mathcal{P}') = \operatorname{Min}(\mathcal{P}'_1) \times \cdots \times \operatorname{Min}(\mathcal{P}'_n)$ and $\operatorname{Min}(\mathcal{P}'_i)$ is the set of minimal elements of $\mathcal{P}'_i$. Now the latter problem is easily seen to be polynomially solvable as follows. Let $\operatorname{Min}(\mathcal{P}'_i) = \{q^1_i, \ldots, q^{k_i}_i\}$, for $i \in [n]$, where $k_i = |\mathcal{P}'_i|$. By construction, only $l \leq k$ of the posets $\mathcal{P}'_i$ satisfy $\mathcal{P}'_i \neq \mathcal{P}_i$. Assume, without loss of generality, that these posets are $\mathcal{P}'_1 \times \cdots \times \mathcal{P}'_l$; then our problem reduces to finding whether $\{q^{i_1}_1\} \times \cdots \times \{q^{i_l}_l\} \times \operatorname{Min}(\mathcal{P}_{l+1}) \times \cdots \times$

$\mathrm{Min}(\mathcal{P}_n) \subseteq \mathcal{A}^+$ for all $(i_1, \ldots, i_l) \in [k_1] \times \cdots \times [k_l]$. Each such problem is equivalent to determining whether $\mathrm{Min}(\mathcal{P}_{l+1}) \times \cdots \times \mathrm{Min}(\mathcal{P}_n) \subseteq (\mathcal{A}^{i_1,\ldots,i_l})^+$, where $\mathcal{A}^{i_1,\ldots,i_l} = \{(a_{l+1}, \ldots, a_n) \mid a \in \mathcal{A},\ \ a_j \preceq q_j^{i_j} \text{ for } j = 1, \ldots, l, \text{ and } a_j^+ \cap \mathrm{Min}(\mathcal{P}_j) \neq \emptyset \text{ for } j = l+1, \ldots, n\}$. Note that $\mathcal{A}^{i_1,\ldots,i_l} \subseteq \mathrm{Min}(\mathcal{P}_{l+1}) \times \cdots \times \mathrm{Min}(\mathcal{P}_n)$ since $\mathcal{A} \subseteq \mathcal{P}$ was assumed, and hence, each subproblem of the form $\mathrm{Min}(\mathcal{P}_{l+1}) \times \cdots \times \mathrm{Min}(\mathcal{P}_n) \subseteq (\mathcal{A}^{i_1,\ldots,i_l})^+$ can be solved in $O(W(\mathcal{P})nm)$ as a special case of Proposition 2 (since each of the posets $\mathrm{Min}(\mathcal{P}_{l+1}), \ldots, \mathrm{Min}(\mathcal{P}_n)$ can be decomposed into independent posets of size 1 each). □

On the negative side, if we do not insist on the condition $\mathcal{A}, \mathcal{B} \subseteq \mathcal{P}$ in Proposition 3, then the problem becomes NP-hard even for $\mathcal{B} = \emptyset$.

PROPOSITION 4. *Given a subposet $\mathcal{Q}$ of a poset $\mathcal{P}$ and a subset $\mathcal{A} \subseteq \mathcal{P}$, it is coNP-complete to decide if $\mathcal{Q} \subseteq \mathcal{A}^+$.*

Proof. We use a polynomial transformation from the satisfiability problem. Let $C = C_1 \wedge \cdots \wedge C_m$ be a conjunctive normal form in n variables $x_1, \ldots, x_n$, and let us consider the poset $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$, where $\mathcal{P}_i = \mathcal{Q}_i \cup \{a_i^1, \ldots, a_i^m\}$, $\mathcal{Q}_i = \{x_i, \overline{x}_i\}$, and where we associate a vector $a^j = (a_1^j, \ldots, a_n^j)$ with each clause C_j, $j = 1, \ldots, m$. The relations in the poset $\mathcal{P}$ are defined as follows: For a literal $l_i \in \mathcal{Q}_i$ and an element $a_i^j \in \mathcal{P}_i \setminus \mathcal{Q}_i$, let $a_i^j \prec l_i$ in $\mathcal{P}_i$ if and only if l_i does not appear in clause C_j in C. Finally, let $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_n$, and $\mathcal{A} = \{a^1, \ldots, a^m\}$. Then $\mathcal{Q} \not\subseteq \mathcal{A}^+$ if and only if C is satisfiable. □

3.4. Projection. As seen above, it is necessary throughout the algorithm to maintain the condition $\mathcal{A}, \mathcal{B} \subseteq \mathcal{P}$, so that when we arrive at the base case, we can apply Proposition 3. Clearly, $\mathcal{A}, \mathcal{B} \subseteq \mathcal{P}$ holds initially, but might not hold after decomposing $\mathcal{P}$. To solve this problem, we project the elements of $\mathcal{A}$ and $\mathcal{B}$ on the poset $\mathcal{Q}$, for each newly created subproblem $\mathrm{DUAL}(\mathcal{Q}, \mathcal{A}, \mathcal{B})$. More precisely, if there are $a \in \mathcal{A}$ and $i \in [n]$ such that $a_i^+ \cap \mathcal{Q}_i \neq \emptyset$, but $a_i \notin \mathcal{Q}_i$, we replace a by the set of elements $\{(x, \overline{a}^i) \mid x \in \mathrm{Min}(a_i^+ \cap \mathcal{Q}_i)\}$, where $\mathrm{Min}(\cdot)$ is the set of minimal elements of $(\cdot)$. Similarly, if there is an element $b \in \mathcal{B}$ and an index $i \in [n]$ such that $b_i^- \cap \mathcal{Q}_i \neq \emptyset$, but $b_i \notin \mathcal{Q}_i$, then we replace b by the set of elements $\{(x, \overline{b}^i) \mid x \in \mathrm{Max}(b_i^- \cap \mathcal{Q}_i)\}$. Note that condition (4) continues to hold after such replacements.

In general, an element of $\mathcal{A}$ or $\mathcal{B}$ may project to a number of elements in $\mathcal{Q}$. Thus performing a large number of projection steps, we may end up with an exponential increase in the sizes of $\mathcal{A}, \mathcal{B}$. However, for certain classes of posets, such as lattices and forests with connected precedence graphs (i.e., trees), each element of $\mathcal{A}, \mathcal{B}$ projects to a single element in $\mathcal{Q}$, i.e., $|\mathrm{Min}(a_i^+ \cap \mathcal{Q}_i)| = |\mathrm{Max}(b_i^- \cap \mathcal{Q}_i)| = 1$ for all $a \in \mathcal{A}$, $b \in \mathcal{B}$, and $i \in [n]$. Indeed, if $\mathcal{Q}_i$ is a lattice, then $\mathrm{Min}(a_i^+ \cap \mathcal{Q}_i) = \{a_i \vee \min(\mathcal{Q}_i)\}$ and $\mathrm{Max}(b_i^- \cap \mathcal{Q}_i) = \{b_i \wedge \max(\mathcal{Q}_i)\}$, where $\min(\mathcal{Q}_i)$ and $\max(\mathcal{Q}_i)$ are, respectively, the minimum and maximum elements of $\mathcal{Q}_i$. Similarly, if the precedence graph of $\mathcal{Q}_i$ is a tree and there are two distinct minimal elements $y, z \in \mathcal{Q}_i$ with the property that $y \succ a_i$ and $z \succ a_i$, then there exists an undirected path between y and z in the precedence graph of $\mathcal{Q}_i$ and another path through a_i, forming a cycle, in contradiction to the fact that the original poset $\mathcal{P}_i$ (of which $\mathcal{Q}_i$ is subposet) is a forest.

Thus, in conclusion, when decomposing a given dualization problem into a number of subproblems, we need to make sure that, in each resulting subproblem DUAL $(\mathcal{Q}, \mathcal{A}, \mathcal{B})$, the poset $\mathcal{Q}$ is still the product of lattices, or the product of forests with connected precedence graphs. In fact, this is the only place where the algorithms described later fail to work for products of general posets.

Procedure LD-A$(\mathcal{Q}, \mathcal{A}, \mathcal{B})$**:**
Input: A sublattice $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_n \subseteq \mathcal{L}$ and two antichains $\mathcal{A}, \mathcal{B} \subseteq \mathcal{L}$
Output: **true** if $\mathcal{Q} \subseteq (\mathcal{A}^+ \cup \mathcal{B}^-)$ and **false** otherwise

1. $\mathcal{A} \leftarrow \mathcal{A}(\mathcal{Q})$, $\mathcal{B} \leftarrow \mathcal{B}(\mathcal{Q})$
2. $\mathcal{A} \leftarrow$ PROJECT$(\mathcal{Q}, \mathcal{A})$, $\mathcal{B} \leftarrow$ PROJECT$(\mathcal{Q}, \mathcal{B})$
3. **if** $\min\{|\mathcal{A}|, |\mathcal{B}|\} < \delta(W)$ **then**
4. **return** POLY-DUAL$(\mathcal{Q}, \mathcal{A}, \mathcal{B})$
5. Find $i \in [n]$ and $z \in \mathcal{Q}_i$ that satisfy Lemma 5 (iii)–(iv); **if** no such elements exist **then**
6. **return false**
7. **if** z satisfies Lemma 5(iii) **then**
8. **return** LD-A$((\mathcal{Q}_i \cap z^-) \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}) \wedge (\bigwedge_{x \in \mathrm{Min}(\mathcal{Q}_i \setminus z^-)}$ LD-A$((\mathcal{Q}_i \cap x^+) \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}))$
9. **else**
10. **return** LD-A$((\mathcal{Q}_i \cap z^+) \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}) \wedge (\bigwedge_{x \in \mathrm{Max}(\mathcal{Q}_i \setminus z^+)}$ LD-A$((\mathcal{Q}_i \cap x^-) \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}))$

FIG. 6. *The first dualization procedure for lattices.*

4. Dualization in products of join semilattices. Let $\mathcal{L} = \mathcal{L}_1 \times \cdots \times \mathcal{L}_n$, where each $\mathcal{L}_i$ is a join semilattice with maximum element u_i, and let $\mathcal{A} \subseteq \mathcal{L}$ and $\mathcal{B} \subseteq \mathcal{I}(\mathcal{A})$.

We begin with the observation that dualization on products of join semilattices can be reduced in polynomial time to dualization on products of lattices. Indeed, for each join semilattice $\mathcal{L}_i$, let us add a minimum element l_i that precedes every element in $\mathcal{L}_i$. Then it is easy to see that the resulting poset $\mathcal{L}_i' \stackrel{\text{def}}{=} \mathcal{L}_i \cup \{l_i\}$ is a lattice. Given $\mathcal{A}, \mathcal{B} \subseteq \mathcal{L}$ satisfying (4), let us obtain a new set $\mathcal{B}' \subseteq \mathcal{L}' \stackrel{\text{def}}{=} \mathcal{L}_1' \times \cdots \times \mathcal{L}_n'$ by extending $\mathcal{B}$ as follows. For each added minimum element l_i, we define a new element $b \in \mathcal{B}'$ by setting $b_i = l_i$, and $b_j = u_j$ for $j \neq i$. Clearly, condition (4) still holds for the pair $(\mathcal{A}, \mathcal{B} \cup \mathcal{B}')$, and $\mathcal{A}^+ \cup \mathcal{B}^- = \mathcal{L}$ if and only if $\mathcal{A}^+ \cup (\mathcal{B} \cup \mathcal{B}')^- = \mathcal{L}'$ by construction. Thus, for the rest of this section, we shall assume, without loss of generality, that each poset $\mathcal{L}_i$ is a lattice.

Before we prove Theorem 1, we show that the simpler (high-frequency based) algorithm of [FK96] can also be generalized for lattices to get a weaker bound than that of Theorem 1 (in fact, with an exponent linear in W, in contrast to the super-quadratic bound in Theorem 1).

4.1. Algorithm A. The first dualization algorithm for lattices is given in Figure 6. In the algorithm, we use $\delta = \delta(W) = \sqrt{(W+3)\log(W+2)}$, where $W = W(\mathcal{L})$. As usual, the algorithm is called initially with $\mathcal{Q} = \mathcal{L}$. In a general step, we check if there is a frequent element $z \in \cup_{i=1}^n \mathcal{Q}_i$, satisfying Lemma 5 (iii)–(iv) (where $\alpha = 2$). If no such z can be found, then a new element in $\mathcal{Q} \setminus (\mathcal{A}^+ \cup \mathcal{B}^-)$ can be obtained as described in the proof of Lemma 4. Otherwise, a decomposition of $\mathcal{Q}_i$ into a set of lattices can be obtained, and the algorithm is called recursively as in steps 8 and 10.

4.1.1. Analysis of algorithm LD-A.

LEMMA 6. *Let $C(v)$ be the total number of recursive calls of procedure* LD-A $(\mathcal{Q}, \mathcal{A}, \mathcal{B})$ *on a problem of size $v = |\mathcal{A}(\mathcal{Q})||\mathcal{B}(\mathcal{Q})| \geq 1$. Then $C(v) \leq R(v) \stackrel{\text{def}}{=} v^{\ln v/\epsilon}$, where $\epsilon = 1/(2W \ln m)$.*

Proof. If $v \geq 1$, but $\min\{|\mathcal{A}|, |\mathcal{B}|\} \leq \delta$, then step 4 implies that $C(v) = 1 \leq R(v)$. Suppose now that the algorithm proceeds to step 8, and let $\mathcal{Q}' = (\mathcal{Q}_i \cap z^-) \times \overline{\mathcal{Q}}$ and $\mathcal{Q}^x = (\mathcal{Q}_i \cap x^+) \times \overline{\mathcal{Q}}$ for $x \in \mathrm{Min}(\mathcal{Q}_i \setminus z^-)$ be the subposets constructed at that step.

Then it follows from Lemma 5(iii) that $|\mathcal{A}(\mathcal{Q}')| \leq |\mathcal{A}| - 1$ and $|\mathcal{B}(\mathcal{Q}')| \geq \epsilon|\mathcal{B}|$, and thus

$$v(\mathcal{A}(\mathcal{Q}'), \mathcal{A}(\mathcal{Q}')) \leq (|\mathcal{A}| - 1)|\mathcal{B}| \leq v - \delta,$$

$$v(\mathcal{A}(\mathcal{Q}^x), \mathcal{A}(\mathcal{Q}^x)) \leq |\mathcal{A}|(1-\epsilon)|\mathcal{B}| = (1-\epsilon)v.$$

Combined with the fact that $|\operatorname{Min}(\mathcal{Q}_i \setminus z^-)| \leq W$, this leads to the recurrence

$$C(v) \leq 1 + W \cdot C((1-\epsilon)v) + C(v - \delta).$$

We get also a similar recurrence if the algorithm proceeds to step 10 of LD-A. To evaluate this recurrence, we first apply it k times to get $C(v) \leq k + kW \cdot C((1-\epsilon)v) + C(v - k\delta)$. Letting $k = \lceil \frac{v\epsilon}{\delta} \rceil$ yields $C(v) \leq (1 + (W+1)(\frac{v\epsilon}{\delta} + 1))C((1-\epsilon)v)$, and hence $C(v) \leq (1 + (W+1)(\frac{v\epsilon}{\delta} + 1))^{\ln v/\epsilon} = (W + 2 + \frac{W+1}{\delta} v\epsilon)^{\ln v/\epsilon}$. Since $\min\{|\mathcal{A}|, |\mathcal{B}|\} \geq \delta$, we have $v \geq \delta^2$ and thus

$$v\left(1 - \frac{W+1}{\delta}\epsilon\right) \geq \delta^2\left(1 - \frac{W+1}{2\delta W \ln(2\delta)}\right) \geq W + 2$$

for all $W \geq 1$, by our selection of $\delta(W)$, implying that $C(v) \leq v^{\ln v/\epsilon}$. □

Since $v \leq m^2$, we get by combining Proposition 3 and Lemma 6 that the running time of the algorithm is $O(m^{8W \log^2 m + 1}(nW)^{\sqrt{(W+3)\log(W+2)}}\mu)$.

4.2. Algorithm B. This algorithm, shown in Figure 7, does not use high-frequency decomposition; any $a \in \mathcal{A}$, $b \in \mathcal{B}$, and $i \in [n]$ such that $a_i \not\preceq b_i$ can be used as explained in section 3.2.2 (see step 5 of the algorithm). The algorithm chooses between Rules $(R1)$, $(R2)$, and $(R2')$ according to the sizes of the relevant subsets of $\mathcal{A}$ and $\mathcal{B}$. More precisely, define $\epsilon(v) = \rho(W)/\chi(v)$, where $v = v(\mathcal{A}, \mathcal{B})$, $\rho(W) \stackrel{\text{def}}{=} \gamma(W)/W = 2W\ln(W+1)$ and $\chi(v)$ is defined to be the unique positive root of the equation

$$\left(\frac{\chi(v)}{\rho(W)}\right)^{\chi(v)} = \frac{v^W}{(1 - e^{-\rho(W)})(\delta^W - 1)},$$

and observe that $\epsilon(v) < 1$ for $v \geq \delta^2$, $\delta \geq 2$. If both $\epsilon_1^{\mathcal{A}} \stackrel{\text{def}}{=} |\mathcal{A}_{\not\preceq}(b_i)|/|\mathcal{A}|$ and $\epsilon_1^{\mathcal{B}} \stackrel{\text{def}}{=} |\mathcal{B}_{\preceq}(b_i)|/|\mathcal{B}|$ are greater than $\epsilon(v)$, then the algorithm uses Rule $(R1)$, but with the further decomposition of $\mathcal{Q}_i \setminus b_i^-$, to ensure the lattice property. Otherwise, if $\epsilon_1^{\mathcal{A}} \leq \epsilon(v)$, then the algorithm uses Rule $(R2')$. If $\epsilon_1^{\mathcal{A}} > \epsilon(v)$, then there exists an element $z \in \mathcal{Q}_i$, such that $|\mathcal{A}_{\succeq}(z) \geq \epsilon_1^{\mathcal{A}}|\mathcal{A}|/|\operatorname{Min}(\mathcal{Q}_i \setminus b_i^-)| > \epsilon(v)|\mathcal{A}|/W$. Then again the algorithm chooses between Rules $(R1)$ and $(R2)$ according to the sizes of the sets $\mathcal{A}_{\succeq}(z)$ and $\mathcal{B}_{\not\preceq}(z)$ (see steps 15–20).

Finally, it remains to remark that all of the decompositions described above result, indeed, in dualization subproblems over lattices.

4.2.1. Analysis of algorithm LD-B. Again, we measure the reduction in the "effective" volume at each recursion level.

Step 7. From the condition $\min\{\epsilon_1^{\mathcal{A}}, \epsilon_1^{\mathcal{B}}\} > \epsilon(v)$ and $|\operatorname{Min}(\mathcal{Q}_i \setminus b_i^-)| \leq W$, we get the recurrence

$$\begin{aligned} C(v) &\leq 1 + C(|\mathcal{A}_{\preceq}(b_i)||\mathcal{B}|) + |\operatorname{Min}(\mathcal{Q}_i \setminus b_i^-)|C(|\mathcal{A}||\mathcal{B}_{\not\preceq}(b_i)|) \\ &\leq 1 + C((1 - \epsilon_1^{\mathcal{A}})v) + W \cdot C((1 - \epsilon_1^{\mathcal{B}})v) \\ &\leq 1 + (W+1)C((1 - \epsilon(v))v). \end{aligned} \tag{13}$$

Procedure LD-B($\mathcal{Q}, \mathcal{A}, \mathcal{B}$)**:**
Input: A sublattice $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_n \subseteq \mathcal{L}$ and two antichains $\mathcal{A}, \mathcal{B} \subseteq \mathcal{L}$
Output: **true** if $\mathcal{Q} \subseteq (\mathcal{A}^+ \cup \mathcal{B}^-)$ and **false** otherwise

1. $\mathcal{A} \leftarrow \mathcal{A}(\mathcal{Q})$, $\mathcal{B} \leftarrow \mathcal{B}(\mathcal{Q})$
2. $\mathcal{A} \leftarrow$ PROJECT($\mathcal{Q}, \mathcal{A}$), $\mathcal{B} \leftarrow$ PROJECT($\mathcal{Q}, \mathcal{B}$)
3. **if** $\min\{|\mathcal{A}|, |\mathcal{B}|\} < \delta = 2$ **then**
4. **return** POLY-DUAL($\mathcal{Q}, \mathcal{A}, \mathcal{B}$)
5. Let $a \in \mathcal{A}$, $b \in \mathcal{B}$, and $i \in [n]$ be such that $a_i \not\preceq b_i$
6. $\epsilon_1^{\mathcal{A}} \leftarrow \frac{|\mathcal{A}_{\not\preceq}(b_i)|}{|\mathcal{A}|}$ and $\epsilon_1^{\mathcal{B}} \leftarrow \frac{|\mathcal{B}_{\preceq}(b_i)|}{|\mathcal{B}|}$
7. **if** $\min\{\epsilon_1^{\mathcal{A}}, \epsilon_1^{\mathcal{B}}\} > \epsilon(v(\mathcal{A}, \mathcal{B}))$ **then**
8. **return** LD-B($(\mathcal{Q}_i \cap b_i^-) \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}) \wedge (\bigwedge_{x \in \mathrm{Min}(\mathcal{Q}_i \setminus b_i^-)}$ LD-B$((\mathcal{Q}_i \cap x^+) \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}))$
9. **if** $\epsilon_1^{\mathcal{A}} \leq \epsilon(v(\mathcal{A}, \mathcal{B}))$ **then**
10. Let $x^1, \ldots, x^k$ be the elements of $\mathcal{Q}_i \setminus b_i^-$ in topologically nondecreasing order
11. $d \leftarrow$ LD-B($(\mathcal{Q}_i \cap b_i^-) \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}$)
12. **return** $d \wedge \big(\bigwedge_{j \in [k]} \bigwedge_{(b(y) \in \mathcal{B}_{\succeq}(y): y \in (x^j)^{\perp})}$ LD-B $\big(\{x^j\} \times [\overline{\mathcal{Q}} \cap \big(\bigwedge_{y \in (x^j)^{\perp}} \overline{b}(y)\big)^-], \mathcal{A}, \mathcal{B}\big)\big)$
13. Let $z \in \mathcal{Q}_i$ be such that $|\mathcal{A}_{\succeq}(z)| > \epsilon(v(\mathcal{A}, \mathcal{B}))|\mathcal{A}|/W$
14. $\epsilon_2^{\mathcal{A}} \leftarrow \frac{|\mathcal{A}_{\succeq}(z)|}{|\mathcal{A}|}$ and $\epsilon_2^{\mathcal{B}} \leftarrow \frac{|\mathcal{B}_{\not\succeq}(z)|}{|\mathcal{B}|}$
15. **if** $\epsilon_2^{\mathcal{B}} > \epsilon(v(\mathcal{A}, \mathcal{B}))$ **then**
16. **return** LD-B($(\mathcal{Q}_i \cap z^+) \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}) \wedge (\bigwedge_{x \in \mathrm{Max}(\mathcal{Q}_i \setminus z^+)}$ LD-B$((\mathcal{Q}_i \cap x^-) \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}))$
17. **else**
18. Let $y^1, \ldots, y^h$ be the elements of $\mathcal{Q}_i \setminus z^+$ in topologically nonincreasing order
19. $d \leftarrow$ LD-B($(\mathcal{Q}_i \cap z^+) \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}$)
20. **return** $d \wedge \big(\bigwedge_{j \in [k]} \bigwedge_{(a(y) \in \mathcal{A}_{\not\preceq}(x): x \in (y^j)^{\top})}$ LD-B $\big(\{y^j\} \times [\overline{\mathcal{Q}} \cap \big(\bigvee_{x \in (y^j)^{\top}} \overline{a}(x)\big)^+], \mathcal{A}, \mathcal{B}\big)\big)$

FIG. 7. *The second dualization procedure for lattices.*

Steps 11-12. From $\epsilon_1^{\mathcal{A}} \leq \epsilon(v)$ and (11), we get the recurrence

$$C(v) \leq 1 + C(|\mathcal{A}_{\preceq}(b_i)||\mathcal{B}|) + \sum_{j=1}^{k} \left(\prod_{y \in (x^j)^{\perp}} |\mathcal{B}_{\succeq}(y)| \right) C(|\mathcal{A}_{=}(x^j)||\mathcal{B}_{\succeq}(x^j)|)$$

$$\leq 1 + C(|\mathcal{A}_{\preceq}(b_i)||\mathcal{B}|) + |\mathcal{B}|^W \sum_{j=1}^{k} C(|\mathcal{A}_{=}(x^j)||\mathcal{B}_{\succeq}(x^j)|)$$

$$\leq 1 + C((1 - \epsilon_1^{\mathcal{A}})v) + |\mathcal{B}|^W C(\epsilon_1^{\mathcal{A}} v)$$

$$\leq 1 + C((1 - \epsilon_1^{\mathcal{A}})v) + \frac{v^W}{\delta^W} C(\epsilon_1^{\mathcal{A}} v)$$

$$\leq C((1 - \epsilon)v) + \frac{v^W}{\delta^W - 1} C(\epsilon v) \quad \text{for some } \epsilon \in (0, \epsilon(v)], \tag{14}$$

where the second inequality follows from the fact that $|(x^j)^{\perp}| \leq W$, the third inequality follows from $\sum_{j=1}^{k} C(|\mathcal{A}_{=}(x^j)||\mathcal{B}_{\succeq}(x^j)|) \leq C(\sum_{j=1}^{k} |\mathcal{A}_{=}(x^j)||\mathcal{B}_{\succeq}(x^j)|) = C(|\mathcal{A}_{\not\preceq}(b_i)||\mathcal{B}_{\succeq}(x^j)|)$ since $\{\mathcal{A}_{=}(x^j) \mid j = 1, \ldots, k\}$ is a partition of $\mathcal{A}_{\not\preceq}(b_i)$ and the function $C(\cdot)$ is assumed to be superadditive, the fourth inequality follows from $|\mathcal{B}|^W \leq v(|\mathcal{A}|, |\mathcal{B}|)^W / \delta^W$, and the last inequality follows from the fact that $v \geq \delta^2$ and $\delta \geq 2$.

Step 16. Since $\epsilon_2^{\mathcal{A}} > \frac{\epsilon(v)}{W}$ by our selection of $z \in \mathcal{Q}_i$, and $\epsilon_2^{\mathcal{B}} > \epsilon(v)$, we get

$$\begin{aligned} C(v) &\le 1 + C(|\mathcal{A}||\mathcal{B}_{\succeq}(z)|) + |\operatorname{Max}(\mathcal{Q}_i \setminus z^+)|C(|\mathcal{A}_{\not\succeq}(z)||\mathcal{B}|) \\ &\le 1 + C((1-\epsilon_2^{\mathcal{B}})v) + W \cdot C((1-\epsilon_2^{\mathcal{A}})v) \\ &\le 1 + C((1-\epsilon(v))v) + W \cdot C\left(\left(1 - \frac{\epsilon(v)}{W}\right)v\right). \end{aligned} \tag{15}$$

Steps 19–20. Symmetric to steps 11–12, we get again (14).

4.2.2. Proof of Theorem 1. We show by induction on $v = v(\mathcal{A}, \mathcal{B})$ that recurrences (13)–(15) imply that $C(v) \le R(v) \stackrel{\text{def}}{=} v^{\chi(v)}$. Since, for $\min\{|\mathcal{A}|, |\mathcal{B}|\} < \delta$, step 3 of the algorithm implies that $C(v) = 1$, we may assume that $\min\{|\mathcal{A}|, |\mathcal{B}|\} \ge \delta$, i.e., $v \ge \delta^2 = 4$.

Let us consider first recurrence (15). Using the induction hypothesis and the monotonicity of $\mathcal{X}(v)$, we obtain

$$\begin{aligned} C(v) &\le 1 + [(1-\epsilon(v))v]^{\chi(v)} + W\left[\left(1 - \frac{\epsilon(v)}{W}\right)v\right]^{\chi(v)} \\ &\le 1 + \left(e^{-\rho(W)} + We^{-\rho(W)/W}\right) v^{\chi(v)} \le v^{\chi(v)}, \end{aligned} \tag{16}$$

since $1 - e^{-\rho(W)} - We^{-\rho(W)/W} \ge 1/2$ for all $W \ge 1$.

Let us next consider (13) and note that the monotonicity of $C(v)$ implies that $C((1-\epsilon(v))v) \le C((1-\frac{\epsilon(v)}{W})v)$, concluding by (16) that $C(v) \le R(v)$ for this case, too.

Let us now consider (14) and apply induction to get

$$C(v) \le [(1-\epsilon)v]^{\chi(v)} + \frac{v^W}{\delta^W - 1}[\epsilon v]^{\chi(v)} = \psi(\epsilon) v^{\chi(v)},$$

where $\psi(\epsilon) \stackrel{\text{def}}{=} (1-\epsilon)^{\chi(v)} + \frac{v^W}{\delta^W - 1}\epsilon^{\chi(v)}$. Since $\psi(\epsilon)$ is convex in ϵ, $\psi(0) = 1$, $\epsilon \le \epsilon(v)$, and

$$\begin{aligned} \psi(\epsilon(v)) &= \left(1 - \frac{\rho(W)}{\chi(v)}\right)^{\chi(v)} + \frac{v^W}{\delta^W - 1}\left(\frac{\rho(W)}{\chi(v)}\right)^{\chi(v)} \\ &\le e^{-\rho(W)} + \frac{v^W}{\delta^W - 1}\left(\frac{\rho(W)}{\chi(v)}\right)^{\chi(v)} = 1. \end{aligned}$$

By the definition of $\chi(v)$, it follows that $\psi(\epsilon) \le 1$, and hence, $C(v) \le v^{\chi(v)}$.

Note that, for $\delta \ge 2$ and $W \ge 1$, we have $(\chi/\rho(W))^{\chi} < 3\,(v/\delta)^W$, and thus,

$$\chi(v) < \frac{W\log(v/\delta) + \log 3}{\log(\chi/\rho(W))} \sim \frac{W\rho(W)\log v}{\log\log v}.$$

As $v(\mathcal{A}, \mathcal{B}) < m^2$, we get $\chi(v) = o(W\rho(W)\log m)$, concluding the proof of the theorem.

5. Dualization in products of forests.

5.1. The algorithm. Let $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$, where the precedence graph of each poset $\mathcal{P}_i$ is a forest, and let $\mathcal{A}, \mathcal{B} \subseteq \mathcal{P}$ be two antichains satisfying (4). The algorithm is shown in Figure 9.

If the precedence graph of $\mathcal{P}_i$ is not connected, for some $i \in [n]$, we decompose the problem into a number of subproblems over posets with connected precedence graphs (steps 3–4 of FD; cf. Proposition 2).

Starting from step 7 of FD, we decompose $\mathcal{Q} \subseteq \mathcal{P}$ by picking $a \in \mathcal{A}$, $b \in \mathcal{B}$, and an $i \in [n]$, such that $a_i \not\preceq b_i$. If in-deg$(\mathcal{Q}_i) \leq$ out-deg$(\mathcal{Q}_i)$, then we set $\mathcal{Q}'_i \leftarrow \mathcal{Q}_i \cap a_i^+$ and $\mathcal{Q}''_i \leftarrow \mathcal{Q}_i \setminus \mathcal{Q}'_i$; otherwise, we set $\mathcal{Q}''_i \leftarrow \mathcal{Q}_i \cap b_i^-$ and $\mathcal{Q}'_i \leftarrow \mathcal{Q}_i \setminus \mathcal{Q}''_i$. In the latter case, we should use the symmetric versions of the decomposition rules used for the former case, and a brief way to describe this is to replace $\mathcal{P}$ by its dual poset $\mathcal{P}^*$ and exchange the roles of $\mathcal{A}$ and $\mathcal{B}$ in these rules (step 9 of FD). Assume, without loss of generality, in the following that the former case holds.

As in the case of lattices, the algorithm uses the effective volume $v = v(\mathcal{A}, \mathcal{B})$ to compute the threshold

$$\epsilon(v) = \frac{1}{\chi(v)}, \qquad \text{where } \chi(v)^{\chi(v)} = v^d, \ v = v(\mathcal{A}, \mathcal{B}).$$

If the minimum of $\epsilon^{\mathcal{A}} \stackrel{\text{def}}{=} |\mathcal{A}_{\succeq}(a_i)|/|\mathcal{A}|$ and $\epsilon^{\mathcal{B}} \stackrel{\text{def}}{=} |\mathcal{B}_{\not\succeq}(a_i)|/|\mathcal{B}|$ is bigger than $\epsilon(v)$, then Rule $(R1)$ is used for decomposition (step 13 of FD). Otherwise, we proceed as follows. Let $\mathcal{Q}_i^e = \{x \in \mathcal{Q}'_i \mid x^\perp \cap \mathcal{Q}''_i \neq \emptyset\}$ be the set of elements in $\mathcal{Q}'_i$ with immediate predecessors in $\mathcal{Q}''_i$ (see Figure 8). Let, for each $x \in \mathcal{Q}_i^e$, $\mathcal{Q}_i(x) = \{y \in \mathcal{Q}''_i \cap x^- : \ y \notin z^-$ for all $z \in \mathcal{Q}_i^e$ with $z \prec x\}$, $\mathcal{A}(x) = \mathcal{A}(\mathcal{Q}_i(x) \times \overline{\mathcal{Q}})$, and $\mathcal{B}(x) = \mathcal{B}(\mathcal{Q}_i(x) \times \overline{\mathcal{Q}})$. Observe that $\mathcal{Q}_i(x)$ and $\mathcal{Q}_i(y)$ are independent posets for $x \neq y$, $x, y \in \mathcal{Q}_i^e$, and that $\mathcal{B}(x) \cap (\mathcal{Q}'_i \times \overline{\mathcal{Q}}) = \mathcal{B}_{\succeq}(x)$ for all $x \in \mathcal{Q}_i^e$, since the precedence graph of $\mathcal{Q}_i$ is a tree.

Further, letting $\mathcal{Q}_i^r = \mathcal{Q}''_i \setminus \big(\bigcup_{x \in \mathcal{Q}_i^e} \mathcal{Q}_i(x)\big)$, we can apply Rule $(R2)$ but stop the decomposition after processing the first layer $\{y \in x^\perp \mid x \in \mathcal{Q}_i^e\}$. This gives the following set of duality testing subproblems (steps 15–17 of FD):

$$\begin{aligned}
&\mathcal{Q}'_i \times \overline{\mathcal{Q}} \subseteq \mathcal{A}^+ \cup (\mathcal{B}_{\succeq}(a_i))^-, \\
&\mathcal{Q}_i^r \times \overline{\mathcal{Q}} \subseteq (\mathcal{A}_{\not\succeq}(a_i))^+ \cup (\mathcal{B}_{\not\succeq}(a_i))^-, \\
&\mathcal{Q}_i(x) \times (\overline{\mathcal{Q}} \cap \overline{a}^+) \subseteq (\mathcal{A}(x))^+ \cup (\mathcal{B}(x) \setminus \mathcal{B}_{\succeq}(a_i))^- \quad \forall a \in \mathcal{A}_{\preceq}(x), \ x \in \mathcal{Q}_i^e.
\end{aligned} \tag{17}$$

To see the last decomposition in (17), fix an $x \in \mathcal{Q}_i^e$, and make use of Lemma 1 by taking $z \leftarrow x$ and $\mathcal{R}''_i \leftarrow \mathcal{Q}_i(x)$.

Finally, if $\epsilon^{\mathcal{A}} \leq \epsilon(v) < \epsilon^{\mathcal{B}}$, then we use Rule $(R2')$; see steps 19–21 of FD.

5.2. Analysis of algorithm FD. As before, we first write the recurrences corresponding to the different recursive calls. By steps 3–4 of FD and Proposition 2, we may assume that the precedence graph of each poset $\mathcal{Q}_i$ is connected.

Step 13. Suppose that the connected components of $\mathcal{Q}''_i$ are $\mathcal{Q}_i^1, \ldots, \mathcal{Q}_i^h$. Then

$$\begin{aligned}
C(v(\mathcal{A}, \mathcal{B})) &\leq 1 + C(|\mathcal{A}||\mathcal{B}_{\succeq}(a_i)|) + \sum_{j=1}^{h} C(|\mathcal{A}(\mathcal{Q}_i^j \times \overline{\mathcal{Q}})||\mathcal{B}|) \\
&\leq 1 + C(|\mathcal{A}||\mathcal{B}_{\succeq}(a_i)|) + C(|\mathcal{A}_{\not\succeq}(a_i)||\mathcal{B}|) \\
&= 1 + C((1 - \epsilon^{\mathcal{B}})v) + C((1 - \epsilon^{\mathcal{A}})v) \leq 1 + 2C((1 - \epsilon(v))v),
\end{aligned} \tag{18}$$

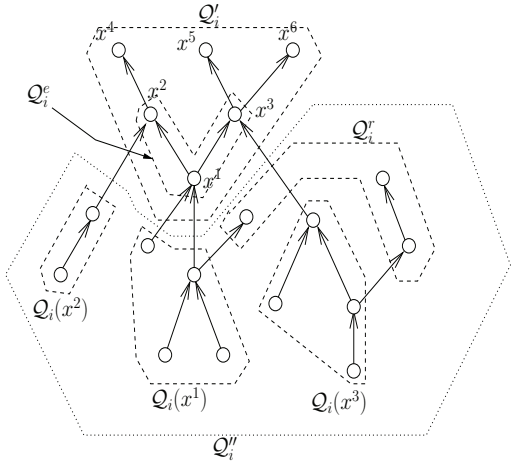


FIG. 8. *Decomposing the forest $\mathcal{Q}_i$.*

Procedure FD($\mathcal{Q}, \mathcal{A}, \mathcal{B}$)**:**
Input: A subposet of a product of forests $\mathcal{Q} = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_n \subseteq \mathcal{P}$ and two antichains $\mathcal{A}, \mathcal{B} \subseteq \mathcal{P}$
Output: **true** if $\mathcal{Q} \subseteq (\mathcal{A}^+ \cup \mathcal{B}^-)$ and **false** otherwise

1. $\mathcal{A} \leftarrow \mathcal{A}(\mathcal{Q})$, $\mathcal{B} \leftarrow \mathcal{B}(\mathcal{Q})$
2. $\mathcal{A} \leftarrow$ PROJECT($\mathcal{Q}, \mathcal{A}$), $\mathcal{B} \leftarrow$ PROJECT($\mathcal{Q}, \mathcal{B}$)
3. **if** there is an $i \in [n]$ such that $\mathcal{Q}_i$ can be decomposed into independent posets $\mathcal{Q}_i^1, \ldots, \mathcal{Q}_i^r$, **then**
4. **return** $\bigwedge_{j=1}^{r}$ FD($\mathcal{Q}_i^j \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}$)
5. **if** $\min\{|\mathcal{A}|, |\mathcal{B}|\} < \delta = 4$ **then**
6. **return** POLY-DUAL($\mathcal{Q}, \mathcal{A}, \mathcal{B}$)
7. Let $a \in \mathcal{A}$, $b \in \mathcal{B}$, and $i \in [n]$ be such that $a_i \not\preceq b_i$
8. **if** in-deg($\mathcal{Q}_i$) > out-deg($\mathcal{Q}_i$) **then**
9. $\mathcal{P} \leftarrow \mathcal{P}^*$, exchange $\mathcal{A}$ and $\mathcal{B}$
10. $\epsilon^{\mathcal{A}} \leftarrow \frac{|\mathcal{A}_{\succeq}(a_i)|}{|\mathcal{A}|}$ and $\epsilon^{\mathcal{B}} \leftarrow \frac{|\mathcal{B}_{\not\succeq}(a_i)|}{|\mathcal{B}|}$
11. Let $\mathcal{Q}_i' \leftarrow \mathcal{Q}_i \cap a_i^+$, $\mathcal{Q}_i'' \leftarrow \mathcal{Q}_i \setminus \mathcal{Q}_i'$
12. **if** $\min\{\epsilon^{\mathcal{A}}, \epsilon^{\mathcal{B}}\} > \epsilon(v(\mathcal{A}, \mathcal{B}))$ **then**
13. **return** FD($\mathcal{Q}_i' \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}$) $\wedge$ FD($\mathcal{Q}_i'' \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}$)
14. **if** $\epsilon^{\mathcal{B}} \leq \epsilon(v(\mathcal{A}, \mathcal{B}))$ **then**
15. Let $\mathcal{Q}_i^e = \{x \in \mathcal{Q}_i' \mid x^{\perp} \cap \mathcal{Q}_i'' \neq \emptyset\}$,
$\mathcal{Q}_i(x) = \{y \in \mathcal{Q}_i'' \cap x^- : \ y \notin z^- \text{ for all } z \in \mathcal{Q}_i^e \text{ with } z \prec x\}$ for $x \in \mathcal{Q}_i^e$, and
$\mathcal{Q}_i^r = \mathcal{Q}_i'' \setminus \left(\bigcup_{x \in \mathcal{Q}_i^e} \mathcal{Q}_i(x)\right)$
16. $d_1 \leftarrow$ FD($\mathcal{Q}_i' \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}$); $d_2 \leftarrow$ FD($\mathcal{Q}_i^r \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}$)
17. **return** $d_1 \wedge d_2 \wedge \left(\bigwedge_{x \in \mathcal{Q}_i^e} \bigwedge_{a \in \mathcal{A}_{\preceq}(x)} \text{FD}(\mathcal{Q}_i(x) \times (\overline{\mathcal{Q}} \cap \overline{a}^+), \mathcal{A}, \mathcal{B})\right)$
18. **else**
19. Let $x^1, \ldots, x^k$ be the elements of $\mathcal{Q}_i'$ in topologically nondecreasing order
20. $d \leftarrow$ FD($\mathcal{Q}_i'' \times \overline{\mathcal{Q}}, \mathcal{A}, \mathcal{B}$)
21. **return** $d \wedge \left(\bigwedge_{j \in [k]} \bigwedge_{(b(y) \in \mathcal{B}_{\geq}(y): y \in (x^j)^{\perp})} \text{FD}\left(\{x^j\} \times \left[\overline{\mathcal{Q}} \cap \left(\bigcap_{y \in (x^j)^{\perp}} \overline{b}(y)^-\right)\right], \mathcal{A}, \mathcal{B}\right)\right)$

FIG. 9. *The dualization procedure for forests.*

since $\mathcal{A}(\mathcal{Q}_i^j \times \overline{\mathcal{Q}})$, for $j = 1, \ldots, h$, partition $\mathcal{A}_{\not\succeq}(a_i)$.

Steps 16–17 *of FD.* From (17) and the fact that $\{\mathcal{A}(x) \mid x \in \mathcal{Q}_i^e\}$ is a partition of $\mathcal{A}$, we get the recurrence

$$
\begin{aligned}
C(v) &\le 1 + C(|\mathcal{A}||\mathcal{B}_{\succeq}(a_i)|) + C(|\mathcal{A}_{\not\succeq}(a_i)||\mathcal{B}_{\not\succeq}(a_i)|) \\
&\quad + \sum_{x \in \mathcal{Q}_i^e} |\mathcal{A}_{\preceq}(x)| C(|\mathcal{A}(x)||\mathcal{B}(x) \setminus \mathcal{B}_{\succeq}(a_i)|) \\
&\le 1 + C((1-\epsilon^{\mathcal{B}})v) + (|\mathcal{A}|+1) C(\epsilon^{\mathcal{B}} v) \\
&\le C((1-\epsilon)v) + \frac{v}{2} C(\epsilon v) \quad \text{for some } \epsilon \in (0, \epsilon(v)],
\end{aligned} \tag{19}
$$

where the last inequality follows from the assumption that $\min\{|\mathcal{A}|, |\mathcal{B}|\} \ge 4$, and hence $|\mathcal{A}| + 1 \le |\mathcal{A}||\mathcal{B}|/3 = v/3$.

Steps 20–21 *of FD.* Since $|(x^j)^{\perp}| \le d$ for every $x^j \in \mathcal{Q}_i'$, by our assumption that in-deg$(\mathcal{Q}_i) \le$ out-deg$(\mathcal{Q}_i)$ (see steps 8–9 of the algorithm), we get

$$
\begin{aligned}
C(v(\mathcal{A}, \mathcal{B})) &\le 1 + C(|\mathcal{A}_{\not\succeq}(a_i)||\mathcal{B}|) + \sum_{j=1}^{k} \left(\prod_{y \in (x^j)^{\perp}} |\mathcal{B}_{\succeq}(y)| \right) C(|\mathcal{A}_{=}(x^j)||\mathcal{B}_{\succeq}(x^j)|) \\
&\le 1 + C(|\mathcal{A}_{\not\succeq}(a_i)||\mathcal{B}|) + |\mathcal{B}|^d C(|\mathcal{A}_{\succeq}(a_i)||\mathcal{B}_{\succeq}(a_i)|) \\
&\le 1 + C((1-\epsilon^{\mathcal{A}})v) + \frac{v^d}{4} C(\epsilon^{\mathcal{A}} v) \\
&\le C((1-\epsilon)v) + \frac{v^d}{2} C(\epsilon v) \quad \text{for some } \epsilon \in (0, \epsilon(v)].
\end{aligned} \tag{20}
$$

Note that this the only place in which the bound d on the degrees appears.

As in subsection 4.2.2, we can show by induction on v that recurrences (18)–(20) imply $C(v) \le R(v) \stackrel{\text{def}}{=} v^{\chi(v)}$. Noting that $\chi(v) < 2\chi(m) \sim 2d \log m / \log\log m$, we get the bound stated in Theorem 2.

6. Dualization algorithm in products of lattices of intervals. Let $\mathcal{L} = \mathcal{L}_1 \times \cdots \times \mathcal{L}_n$ be a product of n lattices of intervals, defined, respectively, by sets of intervals $\mathbb{I}_1, \ldots, \mathbb{I}_n$, and denote by l_i the minimum element of $\mathcal{L}_i$. In this section we prove Theorem 3. We fix $\epsilon = 1/(2 \ln m)$ and use $v(\mathcal{A}, \mathcal{B}, \mathcal{L}) = |\mathcal{A}||\mathcal{B}| \sum_{i=1}^{n} |\mathcal{L}_i|$ as a measure of the volume of the problem.

We begin with the following simple property satisfied by any lattice of intervals.

PROPOSITION 5. *Let $\mathcal{L}_i$ be a lattice of intervals. Then* (i) $|x^{\top}| \le 2$ *for all* $x \ne l_i$ *in* $\mathcal{L}_i$, *and* (ii) $|x^{\perp}| \le 2$ *for all* $x \in \mathcal{L}_i$.

Proof. (i) Assume nonminimum $x \in \mathcal{L}_i$ has $|x^{\top}| \ge 3$. Let I_1, I_2, and I_3 be 3 immediate successors of x in $\mathcal{L}_i$. Let $I_1 = [a, b]$, $I_2 = [c, d]$, where $a, b, c, d \in \mathbb{R}$, and $a < c < b < d$. Then $x = I_1 \cap I_2 = [c, b]$. Let $I_3 = [e, f]$. Now, $I_1 \cap I_3 = x$ implies that $e = c$, and $I_2 \cap I_3 = x$ implies that $f = b$. This gives the contradiction $I_3 = x$.

(ii) Assume $x \in \mathcal{L}_i$ has $|x^{\perp}| \ge 3$. Let I_1, I_2, and I_3 be 3 immediate predecessors of x in $\mathcal{L}_i$. Let $I_1 = [a, b]$, $I_2 = [c, d]$, where $a, b, c, d \in \mathbb{R}$, and $a < b$, $c < d$, $a < c$, and $b < d$. Then $x = Span(I_1, I_2) = [a, d]$. Let $I_3 = [e, f]$. Now, $Span(I_1, I_3) = x$ implies that $f = d$, and $Span(I_2, I_3) = x$ implies that $e = a$. This gives the contradiction $I_3 = x$. ∎

Given subsets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{L}$ that satisfy (4) and a product of lattices of intervals $\mathcal{Q} \subseteq \mathcal{L}$, we follow the general framework as in Figure 4, but use a high-frequency based decomposition. More precisely, assuming $v(\mathcal{A}, \mathcal{B}, \mathcal{Q}) \geq 2$ at a general recursion level, we check if either condition (i) or (ii) of Lemma 5 is satisfied. If neither is satisfied, then we can find an element $x \in \mathcal{Q} \setminus (\mathcal{A}^+ \cup \mathcal{B}^-)$. Otherwise, we consider the following cases.

Case 1. If $i \in [n]$ and $z \in \mathcal{Q}_i$ satisfy condition (i) of Lemma 5 (with $\alpha = 2$), then we consider two subcases.

Case 1.1. If $\mathcal{Q}_i$ is a total order (chain), then use the following decomposition of $\mathcal{Q}_i$: $\mathcal{Q}'_i \leftarrow z^+ \cap \mathcal{Q}_i$, $\mathcal{Q}''_i \leftarrow \mathcal{Q}_i \setminus \mathcal{Q}'_i$. Then $|\mathcal{B}(\mathcal{Q}'_i \times \overline{\mathcal{Q}})| \leq (1-\epsilon)|\mathcal{B}|$ and $|\mathcal{A}(\mathcal{Q}''_i \times \overline{\mathcal{Q}})| \leq |\mathcal{A}| - 1$. This reduces the original problem of volume $v = |\mathcal{A}||\mathcal{B}| \sum_{i=1}^{n} |\mathcal{Q}_i|$ into two subproblems of volumes

$$v' \leq |\mathcal{A}||\mathcal{B}|(1-\epsilon)\left(\sum_{i=1}^{n} |\mathcal{Q}_i| - 1\right) \leq (1-\epsilon)v,$$

$$v'' \leq (|\mathcal{A}| - 1)|\mathcal{B}|\left(\sum_{i=1}^{n} |\mathcal{Q}_i| - 1\right) \leq v - 1.$$

Case 1.2. Otherwise ($\mathcal{Q}_i$ is not a chain), let w be the *largest* element, with respect to the precedence relation on the lattice $\mathcal{Q}_i$, such that $|w^{\perp}| = 2$ (see Figure 10(a)). Denote by q and y, respectively, the two immediate predecessors of w in $\mathcal{Q}_i$, and assume that, without loss of generality, that $|\mathcal{B}_{\preceq}(y)| \geq |\mathcal{B}_{\preceq}(q)|$. It is not hard to see that $\mathcal{Q}_i \cap y^-$ is a lattice of intervals and that $\mathcal{Q}_i \setminus y^-$ is a chain. (Indeed, let $I_q = [a, b]$ and $I_y = [c, d]$ be the two intervals represented by q and y, respectively, and assume that $a < c$ (and therefore $b < d$). Then the former claim follows from the fact that every element in $y^- \subseteq \mathcal{Q}_i$ is associated with an interval, which is the intersection or span of some intervals in $\mathbb{I}_i$, each of which is a subinterval in I_y. The latter claim follows from the fact that if an element $p \in q^- \setminus y^-$ has two immediate predecessors p' and p'' representing intervals $I_{p'} = [e, f]$ and $I_{p''} = [g, h]$, where $e < g$, then we must have $p'' \in y^-$, for otherwise $I_y \subset \mathit{Span}(I_{p''}, I_y) \subset \mathit{Span}(I_p, I_y)$, giving a contradiction with the fact that y is an immediate predecessor of w.)

Now we consider two cases:

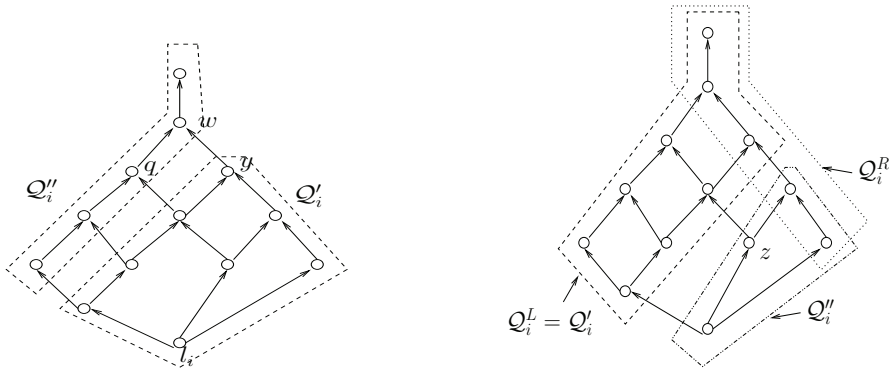
(i) if $z \succ w$, then we use the decomposition $\mathcal{Q}'_i \leftarrow z^+ \cap \mathcal{Q}_i$, $\mathcal{Q}''_i \leftarrow \mathcal{Q}_i \setminus \mathcal{Q}'_i$;
(ii) $z \not\succ w$: in this case, we decompose $\mathcal{Q}_i$ as $\mathcal{Q}'_i \leftarrow \mathcal{Q}_i \cap y^-$, $\mathcal{Q}''_i \leftarrow \mathcal{Q}_i \setminus y^-$.

In case (i), we again get that $|\mathcal{B}(\mathcal{Q}'_i \times \overline{\mathcal{Q}})| \leq (1-\epsilon)|\mathcal{B}|$ and $|\mathcal{A}(\mathcal{Q}''_i \times \overline{\mathcal{Q}})| \leq |\mathcal{A}| - 1$, and consequently, the resulting problems are of respective volumes $v' \leq (1-\epsilon)v$ and $v'' \leq v - 1$. In case (ii), we know that $|\mathcal{B}_{\preceq}(y)| \geq \frac{\epsilon}{2}|\mathcal{B}|$ and thus get $|\mathcal{B}(\mathcal{Q}''_i \times \overline{\mathcal{Q}})| \leq (1 - \epsilon/2)|\mathcal{B}|$ and $|\mathcal{Q}'_i| \leq |\mathcal{Q}_i| - 1$, and therefore, the resulting two problems have volumes $v' \leq v - 1$ and $v'' \leq (1 - \epsilon/2)v$.

Case 2. Now assume that $i \in [n]$ and $z \in \mathcal{Q}_i$ satisfy condition (ii) of Lemma 5. Consider further two subcases.

Case 2.1. If z does not represent the empty interval of $\mathcal{L}_i$, then let $I_z = [a, b]$ be the interval corresponding to z, and let $\mathcal{Q}_i^L \subseteq \mathcal{Q}_i$ be the lattice of intervals $I = [c, d]$ for which $c < a$, and likewise, let $\mathcal{Q}_i^R \subseteq \mathcal{Q}_i$ be the lattice of intervals $I = [e, f]$ for which $f > b$ (see Figure 10(b)). Note that these definitions imply that $(\mathcal{Q}_i^L \cup \{l_i\}) \cap z^- = \{l_i\}$, $(\mathcal{Q}_i^R \cup \{l_i\}) \cap z^- = \{l_i\}$, and $\mathcal{Q}_i^L \cup z^- \cup \mathcal{Q}_i^R = \mathcal{Q}$, where $l_i = \min(\mathcal{Q}_i)$. Note also that $\mathcal{Q}_i^L \cup \mathcal{Q}_i^R \neq \emptyset$ since $z \neq \max(\mathcal{Q}_i)$. By our selection of z, either

$$\text{(i) } |\{a \in \mathcal{A} \mid a_i \in \mathcal{Q}_i^L \setminus \{l_i\}\}| \geq \frac{\epsilon}{2}|\mathcal{A}|, \text{ or (ii) } |\{a \in \mathcal{A} \mid a_i \in \mathcal{Q}_i^R \setminus \{l_i\}\}| \geq \frac{\epsilon}{2}|\mathcal{A}|.$$



(a) Decomposition rule used in Case 1.2 (ii).

(b) Decomposition rule used in Case 2.1.

FIG. 10. *Decomposing the lattice* $\mathcal{L}_i$.

(Note that it is possible that $l_i \in \mathcal{Q}_i^L$ if there are two disjoint intervals in $\mathbb{I}_i$ whose left endpoints are strictly to the left of the left endpoint of z.) In case (i), we decompose $\mathcal{Q}_i$ as follows: $\mathcal{Q}_i' \leftarrow \mathcal{Q}_i^L$, $\mathcal{Q}_i'' \leftarrow (\mathcal{Q}_i \setminus \mathcal{Q}_i') \cup \{l_i\}$. Note that both $\mathcal{Q}_i'$ and $\mathcal{Q}_i''$ are also lattices of intervals, that $|\mathcal{Q}_i'| \leq |\mathcal{Q}_i| - 1$ since $z \notin \mathcal{Q}_i'$, and that $\mathcal{A}(\mathcal{Q}_i'' \times \overline{\mathcal{Q}}) \leq (1-\epsilon/2)|\mathcal{A}|$, since $w \not\preceq y$ for all $w \in \mathcal{Q}_i' \setminus \{l_i\}$ and $y \in \mathcal{Q}_i'' \setminus \{l_i\}$ (indeed, if $I_w = [c,d]$ is the interval corresponding to $w \in \mathcal{Q}_i' \setminus \{l_i\}$ and $I_y = [e,f]$ is the interval corresponding to $y \in \mathcal{Q}_i'' \setminus \{l_i\}$, then $c < a$ while $e \geq a$ and thus $I_w \not\subseteq I_y$). Therefore, we get, in this case, two subproblems of volumes $v' \leq v-1$ and $v'' \leq (1-\epsilon/2)v$. In case (ii), we similarly let $\mathcal{Q}_i' \leftarrow \mathcal{Q}_i^R$ and $\mathcal{Q}_i'' \leftarrow (\mathcal{Q}_i \setminus \mathcal{Q}_i') \cup \{l_i\}$, and we decompose the original problem into two subproblems of volumes $v' \leq v-1$ and $v'' \leq (1-\epsilon/2)v$, respectively.

Case 2.2. Assume now that $z = \min(\mathcal{Q}_i) = l_i$ represents the empty interval of $\mathcal{L}_i$. Note that all immediate successors of z represent pairwise disjoint intervals, and that $|z^\top| \geq 2$. Let z' be the immediate successor of z representing the rightmost such interval $I_{z'} = [a,b]$, and let $\mathcal{Q}_i^L \subseteq \mathcal{Q}_i$ be the lattice of intervals $I = [c,d]$ for which $c < a$. Note in this case that any interval $[c,d]$ in $\mathcal{Q}_i$ either must be strictly to the left of $I_{z'}$, i.e., with $d < a$, or must contain $I_{z'}$ (since z' is the rightmost immediate successor of z). By our choice of z, one of the sets $\{a \in \mathcal{A} \ : \ a_i \in \mathcal{Q}_i^L\}$ or $\mathcal{A}_{\succeq}(z')$ has a size of at least $\frac{\epsilon}{2}|\mathcal{A}|$. In the former case we use the decomposition $\mathcal{Q}_i' \leftarrow \mathcal{Q}_i^L$, $\mathcal{Q}_i'' \leftarrow \mathcal{Q}_i \setminus \mathcal{Q}_i'$, and get two subproblems of volumes $v' \leq v-1$ and $v'' \leq (1-\epsilon/2)v$. In the latter case, we let $\mathcal{Q}_i'' \subseteq \mathcal{Q}_i$ be the lattice of intervals lying strictly to the left of $I_{z'}$ and $\mathcal{Q}_i' \leftarrow ((z')^+ \cap \mathcal{Q}_i) \cup \{z\}$, and get two subproblems of volumes $v' \leq v-1$ and $v'' \leq (1-\epsilon/2)v$.

Thus, in all cases, we apply the algorithm recursively to the resulting subproblems and obtain the recurrence

$$C(v) \leq 1 + C((1-\epsilon/2)v) + C(v-1).$$

Together with $C(v) \leq 1$, for $v \leq 1$, this recurrence evaluates to $C(v) \leq v^{2\log v/\epsilon}$. Since $v \leq m^2 n\mu$, we get that the running time of the algorithm is $O((m^2 n\mu)^{4\ln m \log(m^2 n\mu)})$.

7. Concluding remarks. It is worth mentioning that each poset $\mathcal{P}_i$ belonging to any of the classes of posets considered in this paper has constant dimension; i.e., $\mathcal{P}_i$ is isomorphic to a subposet of the product of a constant number of chains. In

particular, the poset product $\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$, over which we want to solve the dualization problem, can be considered as a subposet of a chain product $\mathcal{C} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_{n'}$, where n' is linear in n. Although we know from [BEG$^+$02] how to solve the dualization problem on products of chains, it is not clear how such a result can be used to solve the original dualization problem on $\mathcal{P}$, since the solution we obtain on $\mathcal{C} \supseteq \mathcal{P}$ (that is, the element $x \in \mathcal{I}(\mathcal{A}) \setminus \mathcal{B}$) might not be an element of $\mathcal{P}$. In fact, as we have seen, the algorithms presented for these classes of posets depend heavily on the type of poset under consideration. This naturally raises the question whether a more general approach can unify these results for posets $\mathcal{P}_i$ of bounded dimension.

It is also not clear whether it is possible to solve the dualization problem in the products of lattices of intervals in time $k^{o(\log k)}$, where $k = |\mathcal{A}| + |\mathcal{B}| + \sum_{i=1}^{n} |\mathcal{L}_i|$, by following a set of decomposition rules, as those used in section 4.2 to solve the problem for general lattices. It seems that if this is to be achieved, then some new decomposition rules are needed, since the current rules in section 4.2 depend exponentially on the maximum out-degree of the lattice $\mathcal{L}_i$, which is $O(|\mathbb{I}_i|)$ in the case of a lattice of a set of intervals $\mathbb{I}_i$.

Finally, we note that, for the more general case of products of arbitrary posets, it remains open whether the problem can be solved in quasi-polynomial time, even for posets $\mathcal{P}_i$ of small size.

Acknowledgments. The author is grateful to Endre Boros, Vladimir Gurvich, Leonid Khachiyan, and Kazuhisa Makino for helpful discussions, and to two anonymous reviewers for useful comments.

REFERENCES

[AB92] M. Anthony and N. Biggs, *Computational Learning Theory: An Introduction*, Cambridge University Press, Cambridge, UK, 1992.

[AIS93] R. Agrawal, T. Imieliński, and A. Swami, *Mining association rules between sets of items in large databases*, in SIGMOD '93: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., ACM, New York, 1993, pp. 207–216.

[AMS$^+$96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, *Fast discovery of association rules*, in Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA, 1996, pp. 307–328.

[BI95] J. C. Bioch and T. Ibaraki, *Complexity of identification and dualization of positive Boolean functions*, Inform. and Comput., 123 (1995), pp. 50–63.

[BEGK04] E. Boros, K. Elbassioni, V. Gurvich, and L. Khachiyan, *Enumerating minimal dicuts and strongly connected subgraphs and related geometric problems*, in Integer Programming and Combinatorial Optimization, Lecture Notes in Comput. Sci. 3064, Springer, Berlin, 2004, pp. 152–162.

[BEG$^+$02] E. Boros, K. Elbassioni, V. Gurvich, L. Khachiyan, and K. Makino, *Dual-bounded generating problems: All minimal integer solutions for a monotone system of linear inequalities*, SIAM J. Comput., 31 (2002), pp. 1624–1643.

[BGKM03] E. Boros, V. Gurvich, L. Khachiyan, and K. Makino, *On maximal frequent and minimal infrequent sets in binary matrices*, Ann. Math. Artif. Intell., 39 (2003), pp. 211–221.

[CDL86] B. Chazelle, R. L. Drysdale, and D. T. Lee, *Computing the largest empty rectangle*, SIAM J. Comput., 15 (1986), pp. 300–315.

[Col87] C. J. Colbourn, *The Combinatorics of Network Reliability*, Oxford University Press, New York, 1987.

[EGLM03] J. Edmonds, J. Gryz, D. Liang, and R. J. Miller, *Mining for empty spaces in large data sets*, Theoret. Comput. Sci., 296 (2003), pp. 435–452.

[EG95] T. Eiter and G. Gottlob, *Identifying the minimal transversals of a hypergraph and related problems*, SIAM J. Comput., 24 (1995), pp. 1278–1304.

[Elb02a] K. Elbassioni, *An algorithm for dualization in products of lattices and its applications*, in ESA, Lecture Notes in Comput. Sci. 2461, Springer, Berlin, 2002, pp. 424–435.

[Elb02b] K. Elbassioni, *On dualization in products of forests*, in STACS, Lecture Notes in Comput. Sci. 2285, Springer, Berlin, 2002, pp. 142–153.

[Elb06] K. Elbassioni, *Finding all minimal infrequent multi-dimensional intervals*, in LATIN 2006, Lecture Notes in Comput. Sci. 3887, Springer, Berlin, 2006, pp. 423–434.

[FK96] M. L. Fredman and L. Khachiyan, *On the complexity of dualization of monotone disjunctive normal forms*, J. Algorithms, 21 (1996), pp. 618–628.

[GMKT97] D. Gunopulos, H. Mannila, R. Khardon, and H. Toivonen, *Data mining, hypergraph transversals, and machine learning (extended abstract)*, in PODS '97: Proceedings of the 16th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Tuscon, AZ, ACM Press, New York, 1997, pp. 209–216.

[Gur75] V. Gurvich, *Nash-solvability of games in pure strategies*, USSR Comput. Math and Math. Phys., 15 (1975), pp. 357–371.

[GK99] V. Gurvich and L. Khachiyan, *On generating the irredundant conjunctive and disjunctive normal forms of monotone Boolean functions*, Discrete Appl. Math., 96/97 (1999), pp. 363–373.

[HCC93] J. Han, Y. Cai, and N. Cercone, *Data-driven discovery of quantitative rules in relational databases*, IEEE Trans. Knowledge Data Engrg., 5 (1993), pp. 29–40.

[HF95] J. Han and Y. Fu, *Discovery of multiple-level association rules from large databases*, in VLDB '95: Proceedings of the 21st International Conference on Very Large Data Bases, Zurich, Switzerland, Morgan Kaufmann, San Francisco, CA, 1995, pp. 420–431.

[KBE+07] L. Khachiyan, E. Boros, K. Elbassioni, V. Gurvich, and K. Makino, *Dual-bounded generating problems: Efficient and inefficient points for discrete probability distributions and sparse boxes for multidimensional data*, Theoret. Comput. Sci., 379 (2007), pp. 361–376.

[LLK80] E. Lawler, J. K. Lenstra, and A. H. G. Rinnooy Kan, *Generating all maximal independent sets: NP-hardness and polynomial-time algorithms*, SIAM J. Comput., 9 (1980), pp. 558–565.

[LKH97] B. Liu, L.-P. Ku, and W. Hsu, *Discovering interesting holes in data*, in Proceedings of the 15th International Conference on Artificial Intelligence (IJCAI), Nagoya, Japan, 1997, pp. 930–935.

[MR95] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, Cambridge, UK, 1995.

[BLQ98] L.-F. Mun, B. Liu, K. Wang, and X.-Z. Qi, *Using decision tree induction for discovering holes in data*, in PRICAI '98: Proceedings of the 5th Pacific Rim International Conference on Artificial Intelligence, Singapore, Springer, London, 1998, pp. 182–193.

[Orl90] M. Orlowski, *A new algorithm for the largest empty rectangle problem*, Algorithmica, 5 (1990), pp. 65–73.

[SA95] R. Srikant and R. Agrawal, *Mining generalized association rules*, in VLDB '95: Proceedings of the 21st International Conference on Very Large Data Bases, Zurich, Switzerland, Morgan Kaufmann, San Francisco, CA, 1995, pp. 407–419.

[SA96] R. Srikant and R. Agrawal, *Mining quantitative association rules in large relational tables*, in SIGMOD '96: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Canada, ACM, New York, 1996, pp. 1–12.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 511–516

LINEAR BOUND ON THE IRREGULARITY STRENGTH AND THE TOTAL VERTEX IRREGULARITY STRENGTH OF GRAPHS*

JAKUB PRZYBYŁO†

Abstract. Let G be a simple graph of order n with no isolated edges and at most one isolated vertex. For a positive integer w, a w-weighting of G is a function $f : E(G) \to \{1, 2, \ldots, w\}$. An irregularity strength of G, $s(G)$, is the smallest w such that there is a w-weighting of G for which $\sum_{e:u\in e} f(e) \neq \sum_{e:v\in e} f(e)$ for all pairs of different vertices $u, v \in V(G)$. We prove that $s(G) < 112\frac{n}{\delta} + 28$, where δ is the minimum degree of G. For d-regular graphs, we strengthen this to $s(G) < 40\frac{n}{d} + 11$. These upper bounds represent improvements of many existing ones. Similar results concerning the "total" version of the irregularity strength are also discussed.

Key words. irregularity strength, total vertex irregularity strength, graph weighting, graph labeling

AMS subject classification. 05C78

DOI. 10.1137/070707385

1. Introduction. All graphs we consider are simple and finite. For a given graph G and its vertex v, $N_G(v)$, $d_G(v)$, $V(G)$, $E(G)$, $\delta(G)$, and $\Delta(G)$ (or simply $N(v)$, $d(v)$, V, E, δ, and Δ) denote the *set of neighbors* and the *degree* of v in G, the *set of vertices*, the *set of edges*, the *minimum degree*, and the *maximum degree* of G, respectively. For a positive integer w, an (edge) *w-weighting* of G is a function $f : E \to \{1, \ldots, w\} = [w]$, while a *total w-weighting* of G is a function $f : V \cup E \to [w]$. We call $f(e)$ and $f(v)$ the *weight of an edge* $e \in E$ and the *weight of a vertex* $v \in V$, respectively, and the greatest value of f is called the *strength* of f. The *induced weight* of $v \in V$ in turn is defined as $c_f(v) = \sum_{u\in N(v)} f(vu)$ if f is an edge weighting or $c_f(v) = f(v) + \sum_{u\in N(v)} f(vu)$ if f is a total weighting of G. We say that a weighting f is *irregular* if the obtained induced weights of all vertices are distinct. The smallest strength of an irregular w-weighting of G is called the *irregularity strength* of G and is denoted by $s(G)$. If it does not exist, we write $s(G) = \infty$. It is easy to see that $s(G) < \infty$ iff G contains no isolated edges and at most one isolated vertex. Analogously, the smallest strength of an irregular total w-weighting of G is called the *total (vertex) irregularity strength* of G and is denoted by $\mathrm{tvs}(G)$. It is easy to see that it is well defined for all graphs.

The irregularity strength was introduced by Chartrand et al. [4] and was motivated by the well-known fact that a simple graph of order at least 2 must contain a pair of vertices with the same degree. On the other hand, a multigraph can be *irregular*, i.e., the degrees of its vertices can all be distinct. Now suppose we want to multiply the edges of a graph G in order to create an irregular multigraph of it. Then $s(G)$ is equal to the smallest maximum multiplicity of an edge in such a multigraph.

Let G be a graph of order n. In [2] Aigner and Triesch proved that $s(G) \leq n - 1$ if G is connected and different from a triangle, and $s(G) \leq n + 1$ otherwise. In [9] Nierhoff refined their method and showed that $s(G) \leq n - 1$ for all graphs with finite irregularity strength, except for K_3. This bound is tight, e.g., for stars. A natural

*Received by the editors November 5, 2007; accepted for publication (in revised form) August 19, 2008; published electronically February 4, 2009. This research was supported by MNiSzW grant N N201 389134.
http://www.siam.org/journals/sidma/23-1/70738.html
†AGH University of Science and Technology, Al. Mickiewicza 30, 30-059 Kraków, Poland (przybylo@wms.mat.agh.edu.pl).

question rises, What happens if the minimum degree of the graph is (much) greater than 1? A simple counting argument (see, e.g., [4]) shows that $s(G) \geq \lceil \frac{n+d-1}{d} \rceil$ for all d-regular graphs, where $d \geq 2$. On the other hand, Faudree and Lehel conjectured that $\frac{n}{d}$ is also "almost" sufficient, i.e., that there exists an absolute constant c such that $s(G) \leq \frac{n}{d} + c$ for all d-regular graphs with $d \geq 2$, but they managed only to show that $s(G) \leq \frac{n}{2} + 9$ holds for them; see [6]. A similar question was formerly posed by Jacobson. See also [8] for a survey by Lehel on this parameter.

Note that if a graph is not regular, but "most" of its vertices are of the same, say minimal, degree δ, then we must expect to have at least about $\frac{n}{\delta}$ weights available to create an irregular weighting. Actually, the same situation exists in the case of the total irregularity strength, especially if δ is large (since then, the single additional weight at the vertex does not change too much). For the d-regular graphs, one can check that $\mathrm{tvs}(G) \geq \lceil \frac{n+d}{d+1} \rceil$; see [3].

The total irregularity strength was introduced quite recently by Bača et al. as a variant of the irregularity strength. In [3] they showed a few simple bounds on this parameter. Among others, $\lceil \frac{n+\delta}{\Delta+1} \rceil \leq \mathrm{tvs}(G) \leq n + \Delta - 2\delta + 1$ holds for all graphs and $\mathrm{tvs}(G) \leq n - 1 - \lfloor \frac{n-2}{\Delta+1} \rfloor$ is true for graphs with no isolated vertices or edges. Moreover, quite obviously, $\mathrm{tvs}(G) \leq s(G)$.

In this paper we prove new upper bounds for $s(G)$ and $\mathrm{tvs}(G)$ which are linear in $\frac{n}{\delta}$ and hence improve the result by Nierhoff for graphs with δ large enough. In the case of regular graphs, the constants in the bounding linear functions are better; consequently we also obtain the improvement of the result by Faudree and Lehel (in most of the cases). In the next section we recall the latest results concerning bounds on $s(G)$ and discuss their consequences for $\mathrm{tvs}(G)$. Our main results are formally stated at the end of that section. In the last part of the paper we present their proofs.

2. Recent results and their consequences. In the paper [7] from 2002, a sizable step forward in the survey on the irregularity strength was made by Frieze et al.

THEOREM 1 (see [7]). *Let G be a graph of order n with no isolated vertices or edges.*

(a) *If $\Delta \leq \lfloor (\frac{n}{\ln n})^{\frac{1}{4}} \rfloor$, then $s(G) \leq 7n(\frac{1}{\delta} + \frac{1}{\Delta})$.*

(b) *If $\lfloor (\frac{n}{\ln n})^{\frac{1}{4}} \rfloor + 1 \leq \Delta \leq \lfloor n^{\frac{1}{2}} \rfloor$, then $s(G) \leq 60\frac{n}{\delta}$.*

(c) *If $\Delta \geq \lfloor n^{\frac{1}{2}} \rfloor + 1$, $\delta \geq \lceil 6 \log n \rceil$, then $s(G) \leq 336(\log n)\frac{n}{\delta}$.*

A similar theorem, but with better constants, holds in the case of the regular graphs.

THEOREM 2 (see [7]). *Let G be a d-regular graph of order n with no isolated vertices or edges.*

(a) *If $d \leq \lfloor (\frac{n}{\ln n})^{\frac{1}{4}} \rfloor$, then $s(G) \leq 10\frac{n}{d} + 1$.*

(b) *If $\lfloor (\frac{n}{\ln n})^{\frac{1}{4}} \rfloor + 1 \leq d \leq \lfloor n^{\frac{1}{2}} \rfloor$, then $s(G) \leq 48\frac{n}{d} + 1$.*

(c) *If $d \geq \lfloor n^{\frac{1}{2}} \rfloor + 1$, then $s(G) \leq 240(\log n)\frac{n}{d} + 1$.*

These very nice results were recently supplemented (and improved in some cases) by Cuckler and Lazebnik.

THEOREM 3 (see [5]). *Let G be a graph of order n with no isolated vertices or edges.*

(a) *If $\delta \geq 10n^{\frac{3}{4}} \log^{\frac{1}{4}} n$, then $s(G) \leq 48\frac{n}{\delta} + 6$.*

(b) *If G is d-regular with $d \geq 10^{\frac{4}{3}} n^{\frac{2}{3}} \log^{\frac{1}{3}} n$, then $s(G) \leq 48\frac{n}{d} + 6$.*

Let g be a w-weighting of a graph G. To prove a simple lemma, which is crucial

in our reasoning on the total irregularity strength, let us define

$$m_g = \max_{X \subseteq V(G)} \{|X| : c_g(u) = c_g(v) \text{ for all } u, v \in X\}.$$

LEMMA 4. *Let g be a w-weighting of a graph G, $w \geq 2$. Then, there exists an irregular total (wm_g)-weighting of G, or an irregular total $((w-1)m_g+1)$-weighting of G if G is a regular graph.*

Proof. For each vertex $v \in V(G)$, denote its weight class as

$$C_v = \{u \in V(G) : c_g(u) = c_g(v)\}.$$

Note that $|C_v| \leq m_g$ for each v. Define a new weighting $f : E(G) \to \{m_g, \ldots, wm_g\}$ by $f(e) = m_g g(e)$. To create a total weighting of it, it is now sufficient to define the values of f on $V(G)$. Note that the weight classes remained the same under f, and that $c_f(u) - c_f(v) = 0$ or $|c_f(u) - c_f(v)| \geq m_g$ for every pair $u, v \in V(G)$. Therefore, for each weight class, say $C = \{v_1, \ldots, v_t\}$ $(t \leq m_g)$, it is sufficient to set $f(v_i) = i$ for $i = 1, \ldots, t$. It is easy to see that such a total weighting is irregular.

Additionally, if G is a d-regular graph, then we can modify f by decreasing the value of $f(e)$ by $m_g - 1$ for each $e \in E(G)$. This way, the strength of the total weighting obtained will be equal to $(w-1)m_g + 1$. Moreover, this weighting will be irregular, since the induced weight of each vertex was decreased by $d(m_g - 1)$. ∎

It was shown in [7], that "a bit" more weights would also suffice in the case of an edge weighting.

LEMMA 5 (see [7]). *Let G be a graph without isolated vertices or edges, and let g be a w-weighting of G. Then, there exists an irregular $((3w+1)m_g)$-weighting of G, or an irregular $((3w-1)m_g+1)$-weighting of G if G is a regular graph.*

Theorems 1 and 2 were thus the consequences of the lemma above and the following probabilistic results; see [7].

LEMMA 6 (see [7]). *Let G be a graph. If $\Delta \leq (\frac{n}{\ln n})^{\frac{1}{4}}$, then there exists $g : E(G) \to \{1, 2\}$ such that $m_g \leq \frac{n}{\delta} + \frac{n}{\Delta}$.*

LEMMA 7 (see [7]). *Let G be a graph. If $\Delta \leq n^{\frac{1}{2}}$, then there exists $g : E(G) \to \{1, 2, 3\}$ such that $m_g \leq 6\frac{n}{\delta}$.*

LEMMA 8 (see [7]). *Let G be a graph. If $n \geq 10$ and $\delta \geq 10 \log n$, then there exists $g : E(G) \to \{1, 2\}$ such that $m_g \leq 48(\log n)\frac{n}{\delta}$.*

If we now take the total irregularity strength into account, then by the three lemmas above and Lemma 4, we immediately obtain the following two theorems.

THEOREM 9. *Let G be a graph of order n with no isolated vertices.*

(a) *If $\Delta \leq \lfloor (\frac{n}{\ln n})^{\frac{1}{4}} \rfloor$, then* $\mathrm{tvs}(G) \leq 2n(\frac{1}{\delta} + \frac{1}{\Delta})$.

(b) *If $\lfloor (\frac{n}{\ln n})^{\frac{1}{4}} \rfloor + 1 \leq \Delta \leq \lfloor n^{\frac{1}{2}} \rfloor$, then* $\mathrm{tvs}(G) \leq 18\frac{n}{\delta}$.

(c) *If $\Delta \geq \lfloor n^{\frac{1}{2}} \rfloor + 1$, $\delta \geq \lceil 10 \log n \rceil$, then* $\mathrm{tvs}(G) \leq 96(\log n)\frac{n}{\delta}$. ∎

THEOREM 10. *Let G be a d-regular graph of order n with no isolated vertices.*

(a) *If $d \leq \lfloor (\frac{n}{\ln n})^{\frac{1}{4}} \rfloor$, then* $\mathrm{tvs}(G) \leq 2\frac{n}{d} + 1$.

(b) *If $\lfloor (\frac{n}{\ln n})^{\frac{1}{4}} \rfloor + 1 \leq d \leq \lfloor n^{\frac{1}{2}} \rfloor$, then* $\mathrm{tvs}(G) \leq 12\frac{n}{d} + 1$.

(c) *If $d \geq \lfloor n^{\frac{1}{2}} \rfloor + 1$, then* $\mathrm{tvs}(G) \leq 48(\log n)\frac{n}{d} + 1$. ∎

Observe, however, that neither of the theorems that we have already mentioned provided the existence of a general linear in $\frac{n}{\delta}$ upper bound on $s(G)$ or $\mathrm{tvs}(G)$, i.e., the one which holds for all ranges of δ, including the cases of regular graphs. Such bounds are presented in the theorems below, which will be proven in the next section.

The proofs are considerably shorter than the ones from [5] and [7], and are not based on the probabilistic method.

THEOREM 11. *Let G be a graph of order n with no isolated vertices or edges. Then $s(G) < 112\frac{n}{\delta} + 28$.*

THEOREM 12 (see [10]). *Let G be a d-regular graph of order n with no isolated vertices or edges. Then $s(G) < 40\frac{n}{d} + 11$.*

THEOREM 13. *Let G be a graph of order n with no isolated vertices. Then* $\mathrm{tvs}(G) < 32\frac{n}{\delta} + 8$.

THEOREM 14. *Let G be a d-regular graph of order n with no isolated vertices. Then* $\mathrm{tvs}(G) < 8\frac{n}{d} + 3$.

Actually, the constants in Theorem 12 can be reduced by a careful construction, which is quite long, down to $16\frac{n}{d} + 6$; see [10] for details. Still, Theorems 1, 2, 3, 9, and 10 give better bounds in some cases. Theorems 11, 12, 13, and 14 are better in general, though.

3. Proof of linear bounds. Given a graph G, a powerful tool for us is the following theorem by Addario-Berry, Dalal, and Reed on the existence of a spanning subgraph of G with degrees from lists consistent with the specified conditions.

THEOREM 15 (see [1]). *Given a graph $G = (V, E)$ and for all $v \in V$, integers a_v^-, a_v^+ such that $a_v^- \le \lfloor \frac{d(v)}{2} \rfloor \le a_v^+ < d(v)$, and*

$$a_v^+ \le \min\left(\frac{d(v) + a_v^-}{2} + 1, 2a_v^- + 3\right), \tag{1}$$

there exists a spanning subgraph H of G such that $d_H(v) \in \{a_v^-, a_v^- + 1, a_v^+, a_v^+ + 1\}$ for every $v \in V$.

COROLLARY 16. *Let $G = (V, E)$ be a graph with $\delta = \delta(G) > 0$, $\Delta = \Delta(G)$, and let $\lambda = \lceil \frac{\delta}{4} \rceil$. There exists a family of sets A_d, $\delta \le d \le \Delta$, (each) of λ consecutive integers such that given any numbers $a_v \in A_{d_G(v)}$ for each $v \in V$, there exists a spanning (containing all the vertices of G) subgraph H of G such that $d_H(v) \in \{a_v, a_v + 1, a_v + \lambda + 1, a_v + \lambda + 2\}$ for every $v \in V$.*

Proof. It is sufficient to prove that for each $d \in \{\delta, \dots, \Delta\}$, there exists a set A_d of λ consecutive integers such that if $d_G(v) = d$ for a vertex $v \in V$, then any $a_v^- \in A_d$ and $a_v^+ := a_v^- + \lambda + 1$ comply with the requirements of Theorem 15. The only exceptions occur for $d = 1$ and $d = 2$ (hence $\lambda = 1$), when $a_v^- = a_v^+ = 0$ and $a_v^- = 0$, $a_v^+ = 1$, respectively, meet the assumptions of Theorem 15. But then $\{a_v^-, a_v^- + 1, a_v^+, a_v^+ + 1\} \subset \{0, 1, 2, 3\}$; hence it is sufficient to take $A_1 = A_2 = \{0\}$.

Assume then now that $v \in V$, $d = d_G(v) \ge 3$ and set $A_d := \{\lceil \frac{d}{2} \rceil - \lambda - 1, \dots, \lceil \frac{d}{2} \rceil - 2\}$. Clearly $|A_d| = \lambda$. Fix any number $a_v \in A_d$. Since $a_v^- := a_v \le \lceil \frac{d}{2} \rceil - 2 \le \lfloor \frac{d}{2} \rfloor$, $a_v^+ := a_v + \lambda + 1 \ge \lfloor \frac{d}{2} \rfloor$ and $a_v^+ \le \lceil \frac{d}{2} \rceil + \lambda - 1 = \lceil \frac{d}{2} \rceil + \lceil \frac{\delta}{4} \rceil - 1 \le \lceil \frac{d}{2} \rceil + \lceil \frac{d}{4} \rceil - 1 < d$; hence it is sufficient to prove (1) for a_v^- and a_v^+. Note then that $a_v^+ = a_v^- + \lambda + 1 \le a_v^- + \lceil \frac{d}{4} \rceil + 1 = a_v^- + (\lceil \frac{d}{4} \rceil - 2) + 3 \le 2a_v^- + 3$ and $a_v^+ = \frac{a_v^-}{2} + \frac{a_v^-}{2} + \lambda + 1 \le \frac{a_v^-}{2} + \frac{a_v^-}{2} + \lceil \frac{d}{4} \rceil + 1 \le \frac{a_v^-}{2} + \frac{1}{2}(\lceil \frac{d}{2} \rceil - 2) + \lceil \frac{d}{4} \rceil + 1 \le \frac{a_v^-}{2} + \frac{d}{2} + 1$; thus (1) holds. ∎

We shall also make use of the following observation.

LEMMA 17. *Let $I_1, \dots, I_k$ be (each) sets of $\lambda > 0$ consecutive integers, and let $S_1, \dots, S_k$ be any finite pairwise disjoint sets. Denote $S = \bigcup_{1 \le i \le k} S_i$ and $n = |S|$. Then, there exists a function $F : S \to \mathbb{Z}$ such that $F(v) \in I_i$ for each $v \in S_i$, $i = 1, \dots, k$, and*

$$\max_{j \in \mathbb{Z}} |\{v \in S : F(v) = j\}| \le \left\lceil \frac{n}{\lambda} \right\rceil.$$

Proof. Let $F : S \to \mathbb{Z}$ be a function such that $F(v) \in I_i$ for each $v \in S_i$, $i = 1, \ldots, k$, and let $C_j = \{v \in S : F(v) = j\}$ for all $j \in \mathbb{Z}$. Denote the *deficiency* of F as

$$d_F = \sum_{j \in \mathbb{Z}} \max \left\{ |C_j| - \left\lceil \frac{n}{\lambda} \right\rceil, 0 \right\}.$$

If $d_F = 0$, then F complies with our requirements. Choose then such an F that minimizes d_F and assume that $d_F > 0$. Then there exist i, u, and j such that $u \in S_i$, $F(u) = j \in I_i$, and $|C_j| > \lceil \frac{n}{\lambda} \rceil$. Note that if there is $j' \in I_i$ such that $|C_{j'}| < \lceil \frac{n}{\lambda} \rceil$, then setting $F(u) = j'$ (instead of j) would decrease the deficiency of F; hence we may assume that $|C_l| \geq \lceil \frac{n}{\lambda} \rceil$ for each $l \in I_i$. But then we have

$$n = \sum_{l \in \mathbb{Z}} |C_l| \geq \sum_{l \in I_i} |C_l| > \lambda \cdot \left\lceil \frac{n}{\lambda} \right\rceil \geq n,$$

a contradiction. ∎

Observe that the result above is the best possible, e.g., if $k = 1$. Given a set $A \subseteq \mathbb{Z}$ and an integer b, let $b + A$ denote the set $\{b + a : a \in A\}$. Note that if we exchanged (shifted) each set I_i from Lemma 17 with (to) $b + I_i$, $i = 1, \ldots, k$, then the function $F + b$ would satisfy the statement of that lemma.

LEMMA 18. *Let G be a graph of order n with $\delta = \delta(G) > 0$. Then there exists a 2-weighting f of G such that*

$$m_f < 16\frac{n}{\delta} + 4.$$

The same conclusion also holds for every d-regular graph with

$$m_f < 8\frac{n}{d} + 2.$$

Proof. Let $G = (V, E)$ be a graph of order n with $\delta = \delta(G) > 0$ and $\Delta = \Delta(G)$, and set $\lambda = \lceil \frac{\delta}{4} \rceil$. Denote $S_d = \{v \in V : d_G(v) = d\}$ and let A_d be the sets from Corollary 16 ($|A_d| = \lambda$), $d = \delta, \ldots, \Delta$. Suppose now that we have chosen some $a_v \in A_d$ for each vertex $v \in V$ of degree d, $\delta \leq d \leq \Delta$. Then, by Corollary 16, there exists a spanning subgraph H of G, such that

$$d_H(v) \in \{a_v, a_v + 1, a_v + \lambda + 1, a_v + \lambda + 2\} \tag{2}$$

for every $v \in V$. Then, if we set $f(e) = 2$ for $e \in E(H)$, and $f(e) = 1$ for the rest of the edges, we will have

$$c_f(v) = d_G(v) + d_H(v) \tag{3}$$

for each $v \in V$. Let $I_d = d + A_d$ for $d = \delta, \ldots, \Delta$. By Lemma 17, there exists a function $F : V \to \mathbb{Z}$ such that $F(v) \in I_d$ if $d_G(v) = d$, and that

$$\max_{j \in \mathbb{Z}} |\{v \in V : F(v) = j\}| \leq \left\lceil \frac{n}{\lambda} \right\rceil. \tag{4}$$

Now for each $v \in V$ of degree d, $\delta \leq d \leq \Delta$, assume that a_v was chosen in such a way that $d + a_v = F(v)$ (it is possible, since $F(v) \in I_d = d + A_d$). By (2), (3) and such a

choice, we have

$$c_f(v) = F(v) \in I_d \quad \text{or} \tag{5}$$

$$c_f(v) = F(v) + 1 \in (1 + I_d) \quad \text{or} \tag{6}$$

$$c_f(v) = F(v) + \lambda + 1 \in ((\lambda + 1) + I_d) \quad \text{or} \tag{7}$$

$$c_f(v) = F(v) + \lambda + 2 \in ((\lambda + 2) + I_d) \tag{8}$$

for each vertex $v \in V$ of degree d, $\delta \le d \le \Delta$.

Since the inequality (4) holds also if we exchange F with $F + 1$, $F + \lambda + 1$, or $F + \lambda + 2$ (and I_d with $1 + I_d$, $(\lambda + 1) + I_d$, or $(\lambda + 2) + I_d$, $d = \delta, \ldots, \Delta$, respectively), then by (5)–(8), the following is true:

$$\max_{j \in \mathbb{Z}} |\{v \in V : c_f(v) = j\}| \le 4 \left\lceil \frac{n}{\lambda} \right\rceil. \tag{9}$$

Note that if G is a d-regular graph, then $F(v) + 1 < F(u) + \lambda + 1$ for every $v, u \in V$. This is because $F(v), F(u) \in I_d = d + A_d$, where A_d is a set of λ consecutive integers. Consequently, we obtain the following improved inequality for the d-regular graphs:

$$\max_{j \in \mathbb{Z}} |\{v \in V : c_f(v) = j\}| \le 2 \left\lceil \frac{n}{\lambda} \right\rceil. \tag{10}$$

Finally, since $\left\lceil \frac{n}{\lambda} \right\rceil \le \left\lceil \frac{4n}{\delta} \right\rceil < \frac{4n}{\delta} + 1$, by (9) and (10) we obtain the thesis. ∎

Proof of Theorems 11 *and* 12. The results follow by Lemmas 18 and 5. ∎

Proof of Theorems 13 *and* 14. The results follow by Lemmas 18 and 4. ∎

REFERENCES

[1] L. Addario-Berry, K. Dalal, and B. A. Reed, *Degree constrained subgraphs*, in Proceedings of GRACO2005, Electron. Notes Discrete Math. 19, Elsevier, Amsterdam, 2005, pp. 257–263.

[2] M. Aigner and E. Triesch, *Irregular assignments of trees and forests*, SIAM J. Discrete Math., 3 (1990), pp. 439–449.

[3] M. Bača, S. Jendrol, M. Miller, and J. Ryan, *On irregular total labellings*, Discrete Math., 307 (2007), pp. 1378–1388.

[4] G. Chartrand, M. S. Jacobson, J. Lehel, O. R. Oellermann, S. Ruiz, and F. Saba, *Irregular networks*, Congr. Numer., 64 (1988), pp. 197–210.

[5] B. Cuckler and F. Lazebnik, *Irregularity strength of dense graphs*, J. Graph Theory, 58 (2008), pp. 299–313.

[6] R. J. Faudree and J. Lehel, *Bound on the Irregularity Strength of Regular Graphs*, Colloq. Math. Soc. Jańos Bolyai 52, North–Holland, Amsterdam, 1988, pp. 247–256.

[7] A. Frieze, R. J. Gould, M. Karoński, and F. Pfender, *On graph irregularity strength*, J. Graph Theory, 41 (2002), pp. 120–137.

[8] J. Lehel, *Facts and quests on degree irregular assignments*, in Graph Theory, Combinatorics, and Applications, Vol. 2, Wiley, New York, 1991, pp. 765–781.

[9] T. Nierhoff, *A tight bound on the irregularity strength of graphs*, SIAM J. Discrete Math., 13 (2000), pp. 313–323.

[10] J. Przybyło, *Irregularity strength of regular graphs*, Electron. J. Combin., 15 (2008), ♯R82.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 517–526

A GENERAL LOWER BOUND FOR POTENTIALLY H-GRAPHIC SEQUENCES*

MICHAEL J. FERRARA† AND JOHN SCHMITT‡

Abstract. We consider a variation of the classical Turán-type extremal problem as introduced by Erdős, Jacobson, and Lehel in [*Graphs realizing the same degree sequences and their respective clique numbers*, in Graph Theory, Combinatorics, and Applications, Vol. 1, Wiley, New York, 1991, pp. 439–449]. Let π be an n-element graphic sequence and $\sigma(\pi)$ be the sum of the terms in π, that is, the degree sum. Let H be a graph. We wish to determine the smallest m such that any n-term graphic sequence π having $\sigma(\pi) \geq m$ has some realization containing H as a subgraph. Denote this value m by $\sigma(H, n)$. For an arbitrarily chosen H, we construct a graphic sequence $\pi^*(H, n)$ such that $\sigma(\pi^*(H, n)) + 2 \leq \sigma(H, n)$. Furthermore, we conjecture that equality holds in general, as this is the case for all choices of H where $\sigma(H, n)$ is currently known. We support this conjecture by examining those graphs that are the complement of triangle-free graphs and showing that the conjecture holds despite the wide variety of structure in this class. We will conclude with a brief discussion of a connection between potentially H-graphic sequences and H-saturated graphs of minimum size.

Key words. degree sequence, potentially graphic sequence, H-saturated graph

AMS subject classifications. Primary, 05C07; Secondary, 05C35

DOI. 10.1137/080715275

1. Introduction. A good reference for any undefined terms is [1]. Let G be a simple undirected graph, without loops or multiple edges. Let $V(G)$ and $E(G)$ denote the vertex set and edge set of G, respectively, and let $d(v)$ denote the degree of a vertex v. Let $\overline{G}$ denote the complement of G. Denote the complete graph on t vertices and the complete bipartite graph with partite sets of size r and s by K_t and $K_{r,s}$, respectively. Additionally, let K_s^t denote the complete balanced multipartite graph with t partite sets of size s. Given any two graphs G and H, their join, denoted $G + H$, is the graph with $V(G + H) = V(G) \cup V(H)$ and $E(G + H) = E(G) \cup E(H) \cup \{gh \mid g \in V(G),\ h \in V(H)\}$. Additionally, let $\alpha(G)$ denote the independence number of G. If H is a subgraph of G, we will write $H \subset G$, and if H is an induced subgraph of G, we will write $H < G$.

A sequence of nonnegative integers $\pi = (d_1, d_2, \ldots, d_n)$ is called *graphic* if there is a (simple) graph G of order n having degree sequence π. In this case, G is said to *realize* π, and we will write $\pi = \pi(G)$. If a sequence π consists of the terms $d_1, \ldots, d_t$ having multiplicities $\mu_1, \ldots, \mu_t$, we may write $\pi = ({d_1}^{\mu_1}, \ldots, {d_t}^{\mu_t})$.

For a given graph H, a sequence π is said to be *potentially H-graphic* if there is some realization of π which contains H as a subgraph. Additionally, let $\sigma(\pi)$ denote the sum of the terms of π. Define $\sigma(H, n)$ to be the smallest integer m so that every n-term graphic sequence π with $\sigma(\pi) \geq m$ is potentially H-graphic. In this paper, given an arbitrary H, we construct a graphic sequence $\pi^*(H, n)$ such that $\sigma(\pi^*(H, n)) + 2 \leq \sigma(H, n)$. We then show that equality holds for all graphs H that are the complement of a triangle-free graph. There have been numerous papers, including

*Received by the editors February 7, 2008; accepted for publication (in revised form) September 25, 2008; published electronically February 4, 2009.
http://www.siam.org/journals/sidma/23-1/71527.html
†Department of Theoretical and Applied Mathematics, The University of Akron, Akron, OH 44325-4002 (mjf@uakron.edu).
‡Department of Mathematics, Middlebury College, Middlebury, VT 05753 (jschmitt@middlebury.edu).

but certainly not limited to [5], [3], [4], [7], [9], [11], [12], [14], [15], [16], [17], [18], [19], and [21], that consider the potential problem for specific graphs or narrow families of graphs. It is our hope that the ideas and results presented in this paper will facilitate a broader consideration of problems of this type.

2. A short history. In this section, we present the extremal sequences for two classes of graphs: complete graphs and complete balanced bipartite graphs. Our goal is to motivate the general constructions in the next section.

2.1. $H = K_t$. In [7] Erdős, Jacobson, and Lehel conjectured that $\sigma(K_t, n) = (t-2)(2n-t+1)+2$. The conjecture arises from consideration of the graph $K_{(t-2)} + \overline{K}_{(n-t+2)}$. It is easy to observe that this graph contains no K_t, is the unique realization of the sequence $((n-1)^{t-2}, (t-2)^{n-t+2})$, and has degree sum $(t-2)(2n-t+1)$. The cases $t = 3, 4$, and 5 were proved separately (see, respectively, [7], [12], [15], and [16]), and Li, Song, and Luo [17] proved the conjecture true via linear algebraic techniques for $t \geq 6$ and $n \geq \binom{t}{2} + 3$. A purely graph-theoretic proof was given in [10] and also as a corollary to the main result in [4].

2.2. $H = K_{s,s}$. The following results appear in [12] and [18]. Here E_1, E_2, E_3, and E_4 are somewhat technical numerical classes which, based on the parity of n and s, ensure that the given degree sums are even.

THEOREM 2.1.

- *If s is an odd, positive integer and $n \geq 4s^2 + 3s - 8$, then*

$$\sigma(K_{s,s}, n) = \begin{cases} (\frac{5}{2}s - \frac{5}{2})n - \frac{11}{8}s^2 + \frac{5}{2}s + \frac{7}{8} & \text{if } (s,n) \in E_3, \\ (\frac{5}{2}s - \frac{5}{2})n - \frac{11}{8}s^2 + \frac{5}{2}s + \frac{15}{8} & \text{if } (s,n) \in E_4. \end{cases} \tag{1}$$

- *If s is an even, positive integer and $n \geq 4s^2 - s - 6$, then*

$$\sigma(K_{s,s}, n) = \begin{cases} (\frac{5}{2}s - 2)n - \frac{11}{8}s^2 + \frac{5}{4}s + 2 & \text{if } (s,n) \in E_1, \\ (\frac{5}{2}s - 2)n - \frac{11}{8}s^2 + \frac{5}{4}s + 1 & \text{if } (s,n) \in E_2. \end{cases} \tag{2}$$

In order to establish a lower bound on $\sigma(K_{s,s}, n)$, the authors present several sequences dependent on the parities of s and n.

(i) If s is odd and $(s,n) \in E_3$, then

$$\pi(K_{s,s}, n) = \left((n-1)^{s-1}, 2s-2, 2s-3, \ldots, \frac{3}{2}s + \frac{3}{2}, \frac{3}{2}s + \frac{1}{2}, \left(\frac{3}{2}s - \frac{1}{2}\right)^{\frac{s}{2}+\frac{3}{2}}, \left(\frac{3}{2}s - \frac{3}{2}\right)^{n-2s}, \frac{3}{2}s - \frac{5}{2}\right). \tag{3}$$

(ii) If s is odd and $(s,n) \in E_4$, then

$$\pi(K_{s,s}, n) = \left((n-1)^{s-1}, 2s-2, 2s-3, \ldots, \frac{3}{2}s + \frac{3}{2}, \frac{3}{2}s + \frac{1}{2}, \left(\frac{3}{2}s - \frac{1}{2}\right)^{\frac{s}{2}+\frac{3}{2}}, \left(\frac{3}{2}s - \frac{3}{2}\right)^{n-2s+1}\right). \tag{4}$$

(iii) If s is even and $(s,n) \in E_1$, then

$$\pi(K_{s,s}, n) = \left((n-1)^{s-1}, 2s-2, 2s-3, \ldots, \frac{3}{2}s + 1, \frac{3}{2}s, \left(\frac{3}{2}s - 1\right)^{n-\frac{3}{2}s+2)}\right). \tag{5}$$

(iv) If s is even and $(s,n) \in E_2$, then

$$\pi(K_{s,s},n) = \Bigg((n-1)^{s-1}, 2s-2, 2s-3, \ldots, \frac{3}{2}s+1, \frac{3}{2}s, \left(\frac{3}{2}s-1\right)^{n-\frac{3}{2}s+1}, \left(\frac{3}{2}s-2\right)\Bigg). \tag{6}$$

Each of these sequences can be realized by the join of K_{s-1} and some graph H'. This H' has no vertices of degree s, one vertex of degree $s-1$, two vertices of degree $s-2$, and so on. More generally, no choice of H' contains x_1 vertices of degree x_2, where $x_1+x_2 = s+1$. This implies that H' cannot possibly contain a copy of K_{x_1,x_2}. However, if any of these sequences were to be potentially $K_{s,s}$-graphic, at least $s+1$ of the vertices in a copy of $K_{s,s}$ would have to be chosen from H'. These vertices, in turn, would comprise some K_{x_1,x_2}, where $x_1+x_2 = s+1$.

3. A general lower bound. We assume that H has no isolated vertices and furthermore that n is sufficiently large relative to $|V(H)|$. We define the quantities

$$u(H) = |V(H)| - \alpha(H) - 1$$

and

$$d(H) = \min\{\Delta(F) : F < H, |V(F)| = \alpha(H)+1\}.$$

Consider the following sequence:

$$\widehat{\pi}(H,n) = ((n-1)^{u(H)}, (u(H)+d(H)-1)^{n-u(H)}). \tag{7}$$

If this sequence is not graphic, that is, if $n-u(H)$ and $d(H)-1$ are both odd, we reduce the smallest term by one. To see that this will result in a graphic sequence, we make two observations. First, $(d(H)-1)$-regular graphs of order $n-u(H) \geq d(H)$ exist whenever $d(H)-1$ and $n-u(H)$ are not both odd. If n and $d(H)-1$ are both odd, it is not difficult to show that the sequence $((d(H)-1)^{n-u(H)-1}, d(H)-2)$ is graphic.

Every realization of $\widehat{\pi}(H,n)$ is a complete graph on $u(H)$ vertices, joined to a graph (call it G') that is either $(d(H)-1)$-regular or nearly so. Note that the subgraph induced by any $\alpha(H)+1$ vertices of H has maximum degree at least $d(H)$. Thus, no realization of $\widehat{\pi}(H,n)$ could possibly contain a copy of H, as at least $\alpha(H)+1$ vertices of such a subgraph would have to lie in G'.

The degree sum of (7) is

$$\sigma(\widehat{\pi}(H,n)) = n(2u(H)+d(H)-1) - u(H)(u(H)+d(H)), \tag{8}$$

and if both $n-u(H)$ and $d(H)-1$ are odd, the sum will be one smaller.

To gain some additional insight, we will consider first the case $H = K_t$. Then $u(K_t) = t-2$ and $d(K_t) = 1$ so that

$$\widehat{\pi}(K_t,n) = ((n-1)^{t-2}, (t-2)^{n-t+2}).$$

This is exactly the extremal sequence put forth to establish the lower bound for $\sigma(K_t,n)$. Similarly, the extremal sequences used to determine $\sigma(kK_2,n), \sigma(C_{2k+1},n)$,

and $\sigma(K_1+kK_2,n)$ are precisely $\widehat{\pi}(kK_2,n)$, $\widehat{\pi}(C_{2k+1},n)$, and $\widehat{\pi}(K_1+kK_2,n)$, respectively (see [12], [14], and [11]). However, $\sigma(\widehat{\pi}(K_{s,s},n))$ is asymptotically equivalent to, but smaller than, $\sigma(K_{s,s},n)$. Along these lines, we are able to refine the sequence given above.

For convenience, let $d=d(H), u=u(H)$, and $\alpha=\alpha(H)$, and let $v_i(H)$ denote the number of vertices of degree i in H. For all $i, d\le i\le\alpha$, we define the quantity m_i to be the minimum number of vertices of degree at least i over all induced subgraphs F of H with $|V(F)|=\alpha+1$ and $\sum_{j=i}^{\alpha}v_j(F)>0$ and 0 if no such subgraphs exist. The quantities n_i, $d\le i\le\alpha$, are defined recursively such that $n_d=m_d-1$ and either $n_i=\min\{m_i-1,n_{i-1}\}$ if $m_i\ge 1$ or $n_i=0$ if $m_i=0$. Finally, we define $\delta_{\alpha-1}=n_{\alpha-1}$, and for $d\le i\le\alpha-2$ we define $\delta_i=n_i-n_{i+1}$. We do not define δ_α, as any induced subgraph composed of a maximum independent set and an additional vertex has at most one vertex of degree α, and as such n_α is always 0.

We now consider the following sequence:

$$\pi^*(H,n)=((n-1)^u,(u+\alpha-1)^{\delta_{\alpha-1}},(u+\alpha-2)^{\delta_{\alpha-2}},\ldots(u+d)^{\delta_d},(u+d-1)^{n-u-\Sigma\delta_i}). \tag{9}$$

The sequence π^* is constructed so that it contains n_i terms that are at least $u+i$ and δ_i terms that are exactly u_i.

If this sequence is not graphic, then we will reduce the smallest term which is strictly greater than $u(H)$ in the sequence by one and redefine $\pi^*(H,n)$ to be this graphic sequence instead. The following is the main result of this paper.

THEOREM 3.1. *Given a graph H, with $u(H)$ and $d(H)$ as above, and n sufficiently large, then*

$$\sigma(H,n)\ge\max\{\sigma(\pi^*(H^*,n))+2\mid H^*\subseteq H\}. \tag{10}$$

Proof. Let H^* be the subgraph of H that realizes the maximum above. Let G be any realization of $\pi^*(H^*,n)$. We show that G does not contain a copy of H^*. Note that this degree sequence implies that G is a copy of $K_{u(H^*)}$ joined to another graph G^* on $n-u(H^*)$ vertices. Assume that there is a copy of H^* contained in G. There are at least $\alpha(H^*)+1$ vertices from G^* that must belong to this copy of H. Let H^{**} denote the subgraph of H^* induced by these $\alpha(H^*)+1$ vertices. Notice, however, that no $\alpha(H^*)+1$ vertices of G^* have sufficient degree to contain a copy of any H^{**}. In particular, if $\sum_{j\ge\ell}v_j(H^{**})>0$, then H^{**} contains at least m_ℓ vertices of degree ℓ or greater. By our construction, there are at most $n_\ell\le m_\ell-1$ vertices of degree at least ℓ in G^*. This contradicts the assumption that $H^{**}\subseteq G^*$. Thus, G contains no copy of H^* and hence no copy of H. ∎

Theorem 3.1 requires that we examine all subgraphs of H. To see that this is necessary, we consider the split graph $K_t+\overline{K_s}$ with a pendant vertex v adjacent to one of the vertices in the independent set of order s. For this choice of H, $\alpha(H)=s$, and hence $u(H)=(s+t+1)-s-1=t$ and $d(H)=1$. However, if we remove v, the pendant vertex, and consider the split graph, we can see that $u(K_t+\overline{K_s})=t-1$ but any $(s+1)$-vertex subgraph of $K_t+\overline{K_s}$ must contain some vertex from the K_t, implying that $d(K_t+\overline{K_s})=s$. Therefore, if we choose $s\ge 3$, $\sigma(\pi^*(K_t+\overline{K_s},n))\ge\sigma(\pi^*(H,n))$.

The reader should note that for any values of n and s, $\pi^*(K_{s,s},n)$ is exactly those sequences given in (3)–(6). Additionally, given values of n,s, and t, $\pi^*(K_s^t,n)$ matches the extremal sequences given in [23].

We conjecture that equality holds in Theorem 3.1.

CONJECTURE 1. *Let H be any graph, and let n be a sufficiently large integer. Then*

$$\sigma(H,n) = \max\{\sigma(\pi^*(H^*,n)) + 2 \mid H^* \subseteq H\}. \tag{11}$$

We also pose the weaker conjecture—that the bound put forth is asymptotically correct.

CONJECTURE 2. *Let H be any graph, and let $\epsilon > 0$. Then there exists an $n_0 = n_0(\epsilon, H)$ such that for any $n > n_0$*

$$\sigma(H,n) \leq \max\{n(2u(H^*) + d(H^*) - 1 + \epsilon) \mid H^* \subseteq H\}. \tag{12}$$

Conjectures 1 and 2 have been verified for a wide variety of graphs. This includes but is not limited to complete graphs and unions of complete graphs [7], [9], [12], [15], [16], [17], complete bipartite graphs [3], [12], [18], complete multipartite graphs [5], [20], matchings [12], cycles [14], (generalized) friendship graphs [2], [9], [11], and split graphs [4]. At this time we know of no subgraph for which these conjectures do not hold for sufficiently large n.

While Conjecture 1 seems challenging, we feel that there is a good chance that Conjecture 2 could be verified. In the following section, we will verify Conjecture 1 for a broad class of graphs.

4. Complements of triangle-free graphs. We now turn our attention to graphs H of order $k \geq 3$ with $\alpha(H) = 2$, or those graphs that are the complement of a triangle-free graph. The main result of this section is as follows.

THEOREM 4.1. *Let H be any graph of order k with $\alpha(H) = 2$. Then*

$$\sigma(H,n) = \sigma(\pi^*(H,n)) + 2.$$

Any graph H in this class has $u(H) = k-3$ and $d(H) \leq 2$. We prove Theorem 4.1 by considering the cases $d(H) = 1$ and $d(H) = 2$ separately. In each case we construct a graph $H(d)$ that contains H as a subgraph and show that $\sigma(H(d),n) = \sigma(\pi^*(H,n))+2$. This implies that $\max\{\sigma(\pi^*(H^*,n))+2 \mid H^* \subseteq H\} = \sigma(\pi^*(H,n))+2$.

The following result from [4] will be very useful.

THEOREM 4.2. *If $n \geq 3s + 2t^2 + 3t - 3$, then*

$$\sigma(K_s + \overline{K}_t, n) = \begin{cases} (t+2s-3)n - (s-1)(s+t-1) + 2 & \text{if } t \text{ or } n-s \text{ is odd,} \\ (t+2s-3)n - (s-1)(s+t-1) + 1 & \text{if } t \text{ and } n-s \text{ are even.} \end{cases}$$

It is not difficult to see that if $d(H) = 2$, then H is isomorphic to $K_k - tK_2$, where k is the order of H and t is some positive integer that is at most $\frac{k}{2}$. Let H be a graph of order $k \geq 3$ with $\alpha(H) = 2$ and $d(H) = 2$, and let $n \geq k$ be an integer. Then, by (9), we have the following.

(i) If $n \equiv k-3 \pmod 2$, then

$$\pi^*(H,n) = ((n-1)^{k-3}, (k-2)^{n-k+3}). \tag{13}$$

(ii) If $n \not\equiv k-3 \pmod 2$, then

$$\pi^*(H,n) = ((n-1)^{k-3}, (k-2)^{n-k+2}, k-3). \tag{14}$$

PROPOSITION 4.3. *Let H be a graph of order k with $\alpha(H) = 2$ and $d(H) = 2$, and let n be a sufficiently large integer. Then*

$$\sigma(H, n) = \sigma(\pi^*(H, n)) + 2 = n(2k - 5) - k^2 + 4k - 1 - m,$$

where $m = n - k + 3 \pmod 2$.

Proof. The fact that $\sigma(H, n) \geq \sigma(\pi^*(H, n)) + 2$ follows from Theorem 3.1. Note that any H with $\alpha(H) = 2$ and $d(H) = 2$ is a subgraph of $K_{k-2} + \overline{K}_2$ so that $\sigma(H, n) \leq \sigma(K_{k-2} + \overline{K}_2, n)$. Theorem 4.2 implies

$$\sigma(K_{k-2} + \overline{K}_2, n) = n(2k - 5) - k^2 + 4k - 1 + m = \sigma(\pi^*(H, n)) + 2.$$

The proposition follows. □

Those graphs H with $\alpha(H) = 2$ and $d(H) = 1$ have a considerably wider variety of structures. Any graph H in this class is the complement of a triangle-free graph G that is not a matching. The disjoint union of two cliques falls into this class, as does $K_k - tP_3$ and many other graphs of varying densities. We are able to verify Conjecture 1 for this diverse class of graphs. Our first observation is that any graph H with $\alpha(H) = 2$ and $d(H) = 1$ must contain $K_2 \cup K_1$ as an induced subgraph, as this is the only graph on three vertices with maximum degree 1. This also immediately implies that $m_d = m_1 = 2$. Therefore, if H is any graph of order k with $\alpha(H) = 2$ and $d(H) = 1$ and $n \geq k$ is an integer, then (9) implies that

$$\pi^*(H, n) = \left((n-1)^{k-3}, (k-3)^{n-k+3}\right). \tag{15}$$

The following lemma from [12] will be useful in the next proof.

LEMMA 4.4. *If π is a graphical sequence with a realization G containing H as a subgraph, then there is a realization G' of π containing H with the vertices of H having the $|V(H)|$ largest degrees of π.*

We now show that Conjecture 1 holds when $\alpha(H) = 2$ and $d(H) = 1$.

PROPOSITION 4.5. *Let H be a graph of order k with $\alpha(H) = 2$ and $d(H) = 1$, and let n be a sufficiently large integer. Then*

$$\sigma(H, n) = \sigma(\pi^*(H, n)) + 2 = n(2k - 6) - k^2 + 5k - 4.$$

Proof. Let π be a nonincreasing, n-term graphic sequence with $\sigma(\pi) \geq n(2k-6) - k^2 + 5k - 4$. Note that if n is sufficiently large, $\sigma(\pi) \geq \sigma(K_{k-1}, n) \geq \sigma(K_{k-3} + \overline{K}_3, n)$. We will show that π has a realization containing $K_{k-3} + (K_2 \cup K_1)$ and, as we have previously observed, that H must contain an induced copy of $K_2 \cup K_1$.

Let G be a realization of π that contains a copy of $K_{k-3} + \overline{K}_3$ on the k vertices of highest degree in G. Such a realization exists by Lemma 4.4. Let S denote this subgraph, let F denote the complete subgraph of order $k - 3$, and let I denote the independent set of order 3 in S so that $S = F + I$. We can assume that F is comprised of the $k - 3$ vertices of highest degree in G. If not, there are vertices x in I and y in F such that $d(y) < d(x)$. We wish to create a realization of G containing a copy of $K_{k-3} + \overline{K}_3$ on the k vertices of highest degree such that x is in F and y is in I. If x is adjacent to all the other vertices in S, we can simply exchange the roles of x and y. If x was not adjacent to exactly one vertex in I, say, v, then as $d(x) > d(y)$ there is some vertex w outside of S that is adjacent to x but not to y. We will create a new realization of π by adding the edges yw and xv and deleting the edges yv and xw. The case where x is not adjacent to exactly two vertices in I is handled

similarly. Repeating this process allows us to create a realization of π containing $K_{k-3} + \overline{K}_3 = F + I$ in which the $k-3$ highest degree vertices of G lie in F.

Let x_1 and x_2 be the vertices in I having the highest degrees, and note that $\sigma(\pi) \geq \sigma(K_{k-1}, n)$ implies $d(x_1)$ and $d(x_2)$ are both at least $k-2$. If there is any edge in the subgraph induced by I, then G contains a copy of $K_{k-3} + (K_2 \cup K_1)$, and we are done. Therefore, we may assume that I is an independent set. Let N_1 and N_2 denote $N(x_1) \setminus S$ and $N(x_2) \setminus S$, respectively, and note that both of these sets are nonempty since $d(x_1)$ and $d(x_2)$ are both at least $k-2$. If y_1 and y_2 are distinct vertices in N_1 and N_2, respectively, then we may assume that y_1 and y_2 are adjacent. If they are not, then we would exchange the edges x_1y_1 and x_2y_2 for the nonedges x_1x_2 and y_1y_2, creating an edge in I and completing the proof.

The goal of the next part of this proof is to show that we may assume that there is some vertex v in F such that $d(v) \leq 4k$.

Consider first the case where $N_2 \subseteq N_1$ ($N_1 \subseteq N_2$ is handled identically), and let w be a vertex in N_2. If $|N_1 \setminus N_2| > k$, then $d(w) > d(x_2)$ since w is adjacent to every vertex in $N_1 \setminus N_2$. We therefore assume that $|N_1 \setminus N_2| \leq k$. Also note that $N_1 \cap N_2$ is a clique and hence contains at most $k-2$ vertices. There is some vertex v in F that is not adjacent to w; otherwise, $d(w) > d(x_1)$, which contradicts our choice of G. Let y be a neighbor of v that does not lie in $S \cup N_1 \cup N_2$. If no such y exists, then clearly $d(v) \leq 4k$. We claim that wy is an edge of G, lest we could exchange the edges x_1w, x_2w, and yv for the nonedges wv, wy, and x_1x_2 (see Figure 1), creating an edge in I. However, if the degree of v is more than $4k$, there are at least $k-1$ such choices for y. This implies that $d(w) \geq k + |N_1| > d(x_1)$, which contradicts our choice of G. Thus we may assume that $d(v) \leq 4k$.

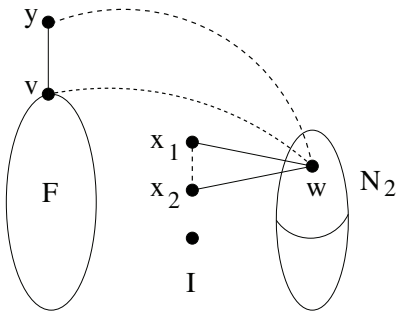


FIG. 1. $N_2 \subseteq N_1$.

Assume now that there is some vertex w_1 in $N_1 \setminus N_2$ and some vertex w_2 in $N_2 \setminus N_1$. We first show that $N_1 \cup N_2$ is complete. To accomplish this, we need only show that for any w_1' in $N_1 \setminus N_2$, w_1w_1' is an edge of G (or, symmetrically, if w_2' is an element of $N_2 \setminus N_1$, then w_2w_2' is an edge in G). If not, we can exchange the edges x_1w_1, x_1w_1', and x_2w_2 for the nonedges w_1w_1', x_1w_2, and x_1x_2, creating an edge in I and completing the proof. Thus, since $N_1 \cup N_2$ is complete, we may assume that $|N_1 \cup N_2| \leq k-1$. Again, there is some v in F such that w_2 is not adjacent to v, lest $d(w_2) > d(x_2)$. Let y be any neighbor of v not in $S \cup N_1 \cup N_2$. Then w_1 is adjacent to y or else we could exchange the edges yv, x_1w_1, and x_2w_2 for the nonedges yw_1, vw_2, and x_1x_2 (see Figure 2), creating an edge in I. If $d(v) > 3k$, then there are at least k such choices for y, implying that $d(w_1) \geq k + |N_1 \cup N_2| - 1 > d(x_1)$, which is a contradiction.

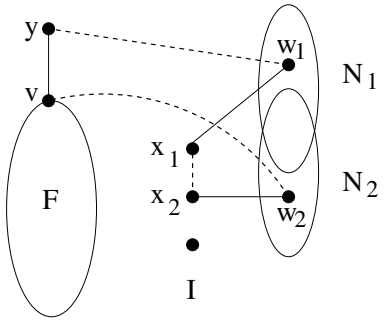


FIG. 2. $N_2 \not\subseteq N_1$ *and* $N_1 \not\subseteq N_2$.

Hence, we may assume that there is some vertex v in F such that $d(v) \leq 4k$. As a result, there are at most $(k-4)(n-1)+4k$ edges adjacent to vertices in F, at most $12k$ edges adjacent to vertices in I, and, as both N_1 and N_2 have at most $4k$ vertices each, at most $4k(8k) = 32k^2$ edges adjacent to vertices in $N_1 \cup N_2$. This is at most $(k-4)n+32k^2+15k+4$ edges. However, there are at least $\sigma(\pi)/2 = (k-3+o(1))n$ edges in G, so for n sufficiently large there is some edge yz in G such that y is not adjacent to any w_1 in N_1 and z is not adjacent to any w_2 in N_2, where w_1 and w_2 may be the same vertex. We can therefore exchange the edges x_1w_1, x_2w_2, and yz for the nonedges w_1y, w_2z, and x_1x_2, creating an edge in I and completing the proof. □

Propositions 4.3 and 4.5 together imply Theorem 4.1. As mentioned above, there is quite a wide variety to the structures of those graphs H having independence number 2, and yet we have demonstrated that $\sigma(H,n)$ for this class depends only on the value of $d(H)$, as suggested by Conjecture 1.

5. *H*-saturated graphs. Here we describe the relationship of $\sigma(H,n)$ to another extremal function $sat(n,H)$. We begin with the relevant terminology and results.

A graph G is said to be H-saturated if G contains no copy of H as a subgraph and for any edge e not in G, $G+e$ does contain a copy of H. The problem of determining the *minimum* number of edges in an H-saturated graph, denoted $sat(n,H)$, was first considered in 1963 by Erdős, Hajnal, and Moon [6] for $H = K_t$. They determined that $sat(n,K_t) = (t-2)(n-1) - \binom{t-2}{2}$, which arises from consideration of the split graph $K_{t-2} + \overline{K}_{n-t+2}$. The best known upper bound for an arbitrary graph H is given by the following result of Kászonyi and Tuza [13].

THEOREM 5.1 (Kászonyi and Tuza [13]). *Let $u(H)$ be as defined above, and set*

$$s(H) = \min\{e(H^*) | \alpha(H^*) = \alpha(H),\ |V(H^*)| = \alpha(H)+1, H^* \subseteq H\};$$

then,

$$sat(n,H) \leq n\left(u(H) + \frac{s(H)-1}{2}\right) - \frac{u(H)\big(u(H)+s(H)\big)}{2}. \tag{16}$$

The reader should note that the bound given in Theorem 5.1 reflects the number of edges in the join of $K_{u(H)}$ and a graph which is (nearly) $(s-1)$-regular. Comparing Theorem 5.1 to the construction of $\pi^*(H,n)$, we note that $d(H) \leq s(H)$ and hence

that if $i \geq s(H)$, $n_i = 0$. Theorems 5.1 and 3.1 immediately imply the following result.

THEOREM 5.2. *Given a graph H, if there exists an $H' \subseteq H$ with $2u(H')+d(H')-1 \geq 2u(H)+s(H)-1$, then for n sufficiently large we have*

$$2sat(n,H) < \sigma(H,n). \tag{17}$$

In particular, this result holds if $d(H) = s(H)$.

We strongly believe that the conclusion of Theorem 5.2 holds in general, even though the hypothesis does not. Therefore, we conjecture the following.

CONJECTURE 3. *Let H be a graph, and let n be a sufficiently large integer. Then*

$$2sat(n,H) < \sigma(H,n).$$

As the problem of determining $sat(n,H)$ has proven difficult over time, we are not able to confirm Conjecture 3 in as many cases as Conjectures 1 and 2. We know that Conjecture 3 holds for complete graphs [6], [7], tK_p, and certain generalized friendship graphs [8], C_4 [12], [22], [24], and $K_{1,t}$ [13].

6. Conclusion. In light of Theorem 4.1, it may be interesting to individually consider classes of graphs with fixed independence numbers. This may be a fruitful direction, although the diversity in the structures of the $(\alpha(H)+1)$ vertex induced subgraphs of such graphs rapidly increases. We feel that this line of investigation would move us closer to the goal of verifying either of Conjectures 1 and 2.

Acknowledgment. The authors would like to thank Mike Jacobson for his helpful comments and insightful questions that led to Theorem 4.1.

REFERENCES

[1] B. BOLLOBÁS, *Extremal Graph Theory*, Academic Press, New York, 1978.
[2] G. CHEN, J. SCHMITT, AND J. H. YIN, *Graphic sequences with a realization containing a generalized friendship graph*, Discrete Math., 308 (2008), pp. 6226–6232.
[3] G. CHEN, J. LI, AND J. YIN, *A variation of a classical Turán-type extremal problem*, European J. Combin., 25 (2004), pp. 989–1002.
[4] G. CHEN AND J. YIN, *On Potentially $K_{r_1,r_2,\ldots,r_m}$-graphic Sequences*, Util. Math., 72 (2007), pp. 149–161.
[5] G. CHEN, M. FERRARA, R. GOULD, AND J. SCHMITT, *Graphic sequences with a realization containing a complete multipartite subgraph*, Discrete Math, 308 (2008), pp. 5712–5721.
[6] P. ERDŐS, A. HAJNAL, AND J. W. MOON, *A problem in graph theory*, Amer. Math. Monthly, 71 (1964), pp. 1107–1110.
[7] P. ERDŐS, M. S. JACOBSON, AND J. LEHEL, *Graphs realizing the same degree sequences and their respective clique numbers*, in Graph Theory, Combinatorics, and Applications, Vol. 1, Alavi, Chartrand, Oellerman, and Schwenk, eds., Wiley, New York, 1991, pp. 439–449.
[8] R. FAUDREE, M. FERRARA, R. GOULD, AND M. JACOBSON, *tK_p-saturated graphs*, Discrete Math., to appear.
[9] M. FERRARA, *Graphic sequences with a realization containing a union of cliques*, Graphs Combin., 23 (2007), pp. 263–269.
[10] M. FERRARA, R. GOULD, AND J. SCHMITT, *Using edge swaps to prove the Erdős–Jacobson–Lehel conjecture*, Bull. Inst. Combin. Appl., to appear.
[11] M. FERRARA, R. GOULD, AND J. SCHMITT, *Graphic sequences with a realization containing a friendship graph*, Ars Combin., 85 (2007), pp. 161–171.
[12] R. GOULD, M. JACOBSON, AND J. LEHEL, *Potentially G-graphic degree sequences*, in Combinatorics, Graph Theory, and Algorithms, Vol. I, Alavi, Lick, and Schwenk, eds., John Wiley & Sons, New York, 1999, pp. 387–400.
[13] L. KÁSZONYI AND Z. TUZA, *Saturated graphs with minimal number of edges*, J. Graph Theory, 10 (1986), pp. 203–210.

[14] C. Lai, *The smallest degree sum that yields potentially C_k-graphical sequences*, J. Combin. Math. Combin. Comput., 49 (2004), pp. 57–64.
[15] J. Li and Z. Song, *An extremal problem on the potentially P_k-graphic sequences*, in Proceedings of the International Symposium on Combinatorics and Applications, W. Y. C. Chen et. al., eds., Tanjin, Nankai University, Nankai, 1996, pp. 269–276.
[16] J. Li and Z. Song, *The smallest degree sum that yields potentially P_k-graphical sequences*, J. Graph Theory, 29 (1998), pp. 63–72.
[17] J. Li, Z. Song, and R. Luo, *The Erdős–Jacobson–Lehel conjecture on potentially P_k-graphic sequences is true*, Sci. China Ser. A, 41 (1998), pp. 510–520.
[18] J. Li and J. Yin, *The smallest degree sum that yields potentially $K_{r,r}$-graphic sequences*, Sci. China Ser. A, 45 (2002), pp. 694–705.
[19] J. Li and J. Yin, *An extremal problem on potentially $K_{r,s}$-graphic sequences*, Discrete Math., 260 (2003), pp. 295–305.
[20] J. Li and J. Yin, *Potentially $K_{r_1,r_2,\ldots,r_l,r,s}$-graphic sequences*, Discrete Math., 307 (2007), pp. 1167–1177.
[21] J. Li and J. Yin, *Two sufficient conditions for a graphic sequence to have a realization with prescribed clique size*, Discrete Math., 301 (2005), pp. 218–227.
[22] L. T. Ollmann, *$K_{2,2}$-saturated graphs with a minimal number of edges*, in Proceedings of the 3rd Southeast Conference on Combinatorics, Graph Theory, and Computing (1972), Utilitas Mathematica Publishing, Winnipeg, Canada, 1973, pp. 367–392.
[23] J. Schmitt, *On Potentially P-graphic Degree Sequences and Saturated Graphs*, Ph.D. Dissertation, Emory University, Atlanta, GA, 2005.
[24] Z. Tuza, *C_4-saturated graphs of minimum size*, Acta Univ. Carolin. Math. Phys., 30 (1989), pp. 161–167.

SIAM J. DISCRETE MATH.
Vol. 23, No. 1, pp. 527–559

DISTRIBUTIVE LATTICES DEFINED FOR REPRESENTATIONS OF RANK TWO SEMISIMPLE LIE ALGEBRAS*

L. WYATT ALVERSON II†, ROBERT G. DONNELLY†, SCOTT J. LEWIS†, MARTI MCCLARD‡, ROBERT PERVINE†, ROBERT A. PROCTOR§, AND N. J. WILDBERGER¶

Abstract. For a rank two root system and a pair of nonnegative integers, using only elementary combinatorics we construct two posets. The constructions are uniform across the root systems $A_1 \oplus A_1$, A_2, C_2, and G_2. Examples appear in Figures 3.2 and 3.3. We then form the distributive lattices of order ideals of these posets. Corollary 5.4 gives elegant quotient-of-products expressions for the rank generating functions of these lattices (thereby providing answers to a 1979 question of Stanley). Also, Theorem 5.3 describes how these lattices provide a new combinatorial setting for the Weyl characters of representations of rank two semisimple Lie algebras. Most of these lattices are new; the rest of them (or related structures) have arisen in the work of Stanley, Kashiwara, Nakashima, Littelmann, and Molev. In a future paper, one author shows that the posets constructed here form a Dynkin diagram-indexed answer to a combinatorially posed classification question. In a companion paper, some of these lattices are used to explicitly construct some representations of rank two semisimple Lie algebras. This implies that these lattices are strongly Sperner.

Key words. distributive lattice, rank generating function, rank two semisimple Lie algebra, representation

AMS subject classifications. 05A15, 05E10, 17B10

DOI. 10.1137/070689887

1. Introduction. One of the earliest combinatorial forays into Lie representation theory was Stanley's [Sta1] in 1979. Certain polynomials arising from representations which had elegant quotient-of-product forms captured his attention. He observed that some of these polynomials were the rank generating functions of certain distributive lattices. In Problem 3 of [Sta1] he asked if further distributive lattices could be found which would be associated to more of the polynomials. Consider the poset "$\mathbf{2} \times \mathbf{3}$" shown in Figure 1.1, the product of chains with two and three elements. Its lattice $L(2,3) = J(\mathbf{2} \times \mathbf{3})$ of order ideals is shown in Figure 1.1. Stanley knew that the rank generating function for the general case $L(k, n+1-k) = J\big(\mathbf{k} \times (\mathbf{n} + \mathbf{1} - \mathbf{k})\big)$ satisfies the identity

$$\sum N_j q^j = \frac{(1-q^{n+1})(1-q^n)\cdots(1-q^{n+2-k})}{(1-q^k)(1-q^{k-1})\cdots(1-q)},$$

where N_j is the number of order ideals in $\mathbf{k} \times (\mathbf{n} + \mathbf{1} - \mathbf{k})$ with j elements. The right-hand side is the "Gaussian coefficient" q-analogue of the binomial coefficient $\binom{n+1}{k}$. It is also a shifted version of the principal specialization of the Weyl character for the

*Received by the editors April 28, 2007; accepted for publication (in revised form) September 26, 2008; published electronically February 4, 2009.
http://www.siam.org/journals/sidma/23-1/68988.html

†Department of Mathematics and Statistics, Murray State University, Murray, KY 42071 (leslie.alverson@murraystate.edu, rob.donnelly@murraystate.edu, scott.lewis@murraystate.edu, bob.pervine@murraystate.edu).

‡Department of Mathematics, University of Tennessee, Knoxville, TN 37996 (mmcclard@math.utk.edu).

§Department of Mathematics, University of North Carolina, Chapel Hill, NC 27599 (rap@email.unc.edu).

¶School of Mathematics, University of New South Wales, Sydney, NSW 2052, Australia (n.wildberger@unsw.edu.au).

kth fundamental representation of the Lie algebra $\mathfrak{sl}(n+1,\mathbb{C})$, the rank n simple Lie algebra of type A. These considerations led Stanley to introduce the more general distributive lattices $L(\lambda, n+1)$, whose elements are semistandard tableaux of shape λ with entries from $\{1,2,\ldots,n+1\}$. Similar identities hold for the rank generating functions of these lattices. Stanley was aware that the polynomial associated to the "last" fundamental representation of the Lie algebra $\mathfrak{sp}(2n,\mathbb{C})$ specializes to the $(n+1)$st Catalan number $\frac{2}{n+2}\binom{2n+1}{n}$ when q is set to 1. Thus the principal specialization of the Weyl character for that representation is a q-analogue of the $(n+1)$st Catalan number. The second author of this paper constructed a poset P_n such that the distributive lattice $L_n = J(P_n)$ of its order ideals has rank generating function $\frac{1-q^2}{1-q^{n+2}}\binom{2n+1}{n}_q$, a shifted version of the principal specialization. So the total number of order ideals from P_n is the $(n+1)$st Catalan number. This result now appears as part (ccc) of Exercise 6.19 of [Sta3]. See Figure 1.1 for the poset P_3; it has 14 order ideals.

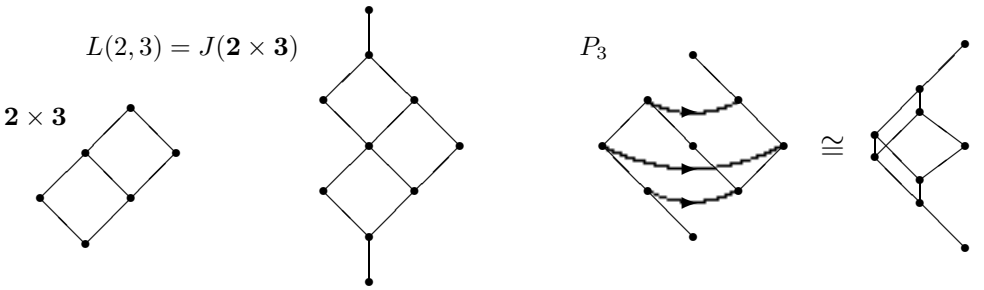


FIG. 1.1. *The distributive lattices of order ideals of the posets* $\mathbf{2}\times\mathbf{3}$ *and* P_3 *answer Stanley's* 1979 *question for certain polynomials.*

Here is the motivating question from [Sta1] posed by Stanley in 1979: "Problem 3: Which other of the polynomials [of Theorem 1] are the rank generating functions for distributive lattices (or perhaps just posets) "naturally associated" with the root system R?" We supply answers to this question by constructing eight two-parameter families of distributive lattices. By the proof of Corollary 5.4, their rank generating functions are the shifted principal specializations of the Weyl characters of the irreducible finite dimensional representations of the four rank two semisimple Lie algebras $A_1 \oplus A_1$, A_2, C_2, and G_2. The answers for C_2 and G_2 are largely new. Given a rank two semisimple Lie algebra $\mathfrak{g}$ and a pair of nonnegative integers, we first construct two "$\mathfrak{g}$-semistandard posets." The "$\mathfrak{g}$-semistandard" distributive lattices are then obtained by ordering the order ideals of these posets by inclusion. For example, the choices of G_2 and nonnegative integer parameters $(2,2)$ specify the last poset in each of Figures 3.2 and 3.3. According to Corollary 5.4, both of these posets have $\frac{1}{5!}(3\cdot 3\cdot 6\cdot 9\cdot 12\cdot 15) = 729 = 3^6$ order ideals. The rank generating function for both of the corresponding G_2-semistandard lattices is

$$RGF_{G_2}(2,2,q) = \frac{(1-q^3)(1-q^3)(1-q^6)(1-q^9)(1-q^{12})(1-q^{15})}{(1-q)(1-q)(1-q^2)(1-q^3)(1-q^4)(1-q^5)}.$$

Hence our lattices $L_{G_2}^{\beta\alpha}(2,2)$ and $L_{G_2}^{\alpha\beta}(2,2)$ are two answers to Problem 3 of [Sta1].

Since the 1970s, the "zoo" of finite sets of combinatorial objects which are enumerated by quotient-of-products formulas has grown to include dozens of species. Here

Corollary 5.4 adds $L_{C_2}^{\beta\alpha}(a,b)$, $L_{C_2}^{\alpha\beta}(a,b)$, $L_{G_2}^{\beta\alpha}(a,b)$, and $L_{G_2}^{\alpha\beta}(a,b)$ to this zoo; they are analogues to the lattices $L(\lambda,n)$. Our $\mathfrak{g}$-semistandard lattices are uniformly defined across the four types of rank two semisimple Lie algebras. Corollary 5.4 also notes that the sequence of rank cardinalities for any $\mathfrak{g}$-semistandard lattice is symmetric and unimodal.

Only familiarity with the most basic Lie representation theory in [Hum] is needed to read this paper. The central fact needed is that each irreducible finite dimensional representation of a semisimple Lie algebra of rank n has a unique n-variate Weyl character.

Some of the rank two $\mathfrak{g}$-semistandard lattices constructed here (or related objects) have appeared in the work of Stanley, Kashiwara, Nakashima, Littelmann, Molev, and several of the authors. However, taken as a whole, each of the C_2- and G_2-families of $\mathfrak{g}$-semistandard lattices is new. The A_2-family of semistandard lattices here is the $n=2$ case of the $L(\lambda,n+1)$ lattices introduced in [Sta1]. A certain infinite subfamily of the C_2-semistandard lattices appeared in [DLP] as the $n=2$ case of the "Molev lattices" $L_B^{Mol}(k,2n)$. A certain infinite subfamily of the G_2-semistandard lattices was studied in [DLP].

Let $\mathfrak{g}$ be a semisimple Lie algebra of rank n. Various data and structures have been associated to each irreducible finite dimensional representation of $\mathfrak{g}$, starting with its highest weight and dimension. Once certain subalgebras of $\mathfrak{g}$ have been fixed, the multiset of weights of a representation is determined. The Weyl character of the representation is the generating function for this multiset of weights. It is a Laurent polynomial in n variables. The polynomials that caught Stanley's eye were shifted versions of the "principal specializations" of the Weyl characters to the variable q. A finer version of Stanley's 1979 question is as follows: For each Weyl character, find a distributive lattice with weighted vertices such that the sum of these weights is the Weyl character. If the lattice elements are assigned weights in a reasonable manner, then a shifted version of the principal specialization will be the lattice's rank generating function. An explicit combinatorial answer to this question (such as a lattice constructed from tableaux) will include a solution to the "labeling problem" for the character: the lattice elements will be combinatorial objects which can be used as labels for the weights. The problem considered here is a stronger version of this finer version of Stanley's question for $n=2$. The "stronger" aspect is described below.

Going further, fixing Chevalley generators for $\mathfrak{g}$ and basis vectors for the representation space determines the data consisting of the entries of the representing matrices for the generators. At this point in several papers (such as [Don1]), the second author introduces the "supporting graph" combinatorial structure. This is a directed graph whose edges are colored by the simple roots of $\mathfrak{g}$. The edges colored by simple root α_i indicate which basis vectors arise with nonzero coefficients when the Chevalley generators x_i and y_i of $\mathfrak{g}$ act on the various basis vectors. This graph is actually the Hasse diagram of a poset. Several of the authors have been able to find distributive lattice supporting graphs for many representations [Don1], [DLP], [ADLP]. The crystal graph is another combinatorial structure associated to a representation. For irreducible representations, the crystal graph is a supporting graph when the weight multiplicities are all one. Such representations have only one supporting graph. But otherwise the crystal graph has fewer edges than do the most efficient supporting graphs; then it cannot support its representation.

Our original goal for developing $\mathfrak{g}$-semisimple lattices was to supply uniformly constructed labels and supporting graphs for explicit realizations of all irreducible

representations of any rank two semisimple Lie algebra $\mathfrak{g}$. Suppose a vertex-weighted edge-colored directed graph is proposed to be a supporting graph of a representation of $\mathfrak{g}$: In addition to its vertex weighting agreeing with the Weyl character, its edge coloring must also satisfy certain conditions specified by the Cartan matrix of $\mathfrak{g}$. (But these conditions alone are not sufficient for the graph to be a supporting graph.) If these edge-coloring conditions are also met, the proposed graph is said to be a "splitting poset" for the representation. The edge-coloring conditions are the embodiment of Stanley's request that the lattices be *natural* with respect to the Lie theory. Here is the "stronger" aspect of our main problem: Not only do we require that the weighting of their elements agree with a given Weyl character, but we also seek edge-colored distributive lattices which are splitting posets. Our answer to this question consists of the $\mathfrak{g}$-semistandard lattices: Proposition 4.2 verifies that the edge colorings satisfy the necessary conditions, and our main result Theorem 5.3 verifies that the vertex weights agree with the character. (The latter verification implies that the order ideals in the $\mathfrak{g}$-semistandard posets can serve as new weight labels for these representations.)

The necessary edge-coloring conditions correspond to the Serre relations (S3) of Proposition 18.1 of [Hum]; the relations (S1) are also satisfied by any splitting poset. Given a splitting poset for a representation of $\mathfrak{g}$, if edge coefficients for the actions of the generators x_i and y_i of $\mathfrak{g}$ can be found that satisfy the relations (S2), then a result of Kashiwara implies that the remaining Serre relations (S_{ij}^{+}) and (S_{ij}^{-}) are automatically satisfied. In certain cases the companion paper [ADLP] is able to attain our original goal by assigning coefficients satisfying (S2) to the edges of the lattices introduced here. So [ADLP] presents explicit realizations for the following irreducible representations of rank two simple Lie algebras, indexed by their type and highest weights: $A_2(a\omega_1 + b\omega_2)$, $C_2(a\omega_1)$, $C_2(b\omega_2)$, $C_2(\omega_1 + b\omega_2)$, $G_2(a\omega_1)$, $G_2(\omega_2)$ for $a, b \geq 0$. Since the $\mathfrak{g}$-semistandard lattices are supporting graphs here, as in [Pr2] they can be seen to be "strongly Sperner." The results of this paper facilitated the new $C_2(\omega_1 + b\omega_2)$ constructions and made it possible to now present the supporting lattices for all of these representations in a uniform fashion.

It can be shown that the $\mathfrak{g}$-semistandard lattices corresponding to the other rank two irreducible representations cannot support their corresponding representations. But to state Corollary 5.4, one needs to know only that the lattice at hand is a splitting poset for an irreducible representation. Hence the beautiful product identities may be written down for the rank generating functions of all $\mathfrak{g}$-semistandard lattices. The necessary edge-coloring conditions are so strong that the second author has been able to prove that the Dynkin diagram-indexed $\mathfrak{g}$-semistandard lattices constitute the entire answer to a purely combinatorial problem [Don2]. See Theorems 6.1 and 6.2.

The positioning of splitting posets (in general; $\mathfrak{g}$-semistandard lattices in particular) in the world of combinatorial structures associated to representations is vaguely similar in spirit to the positioning of crystal graphs: both the lattices and crystal graphs superimpose additional combinatorial structure onto the data contained in the Weyl character, but neither can always support the actions of the corresponding representations. In section 6 we indicate how some splitting posets may hopefully someday be used instead of crystal graphs for some purposes, such as computing tensor products.

Many of the definitions, lemmas, and propositions developed in this paper are needed in [ADLP]. Some of them will also be used in [DW] to explicitly construct many families of splitting posets for the simple Lie algebras A_n, B_n, C_n, D_n, E_6, E_7, and G_2.

Section 2 presents definitions and some preliminary and background results. The reader should initially browse this section and then consult it as needed. Section 3 further considers "grid posets" which were introduced in [ADLP] and whose definition is purely combinatorial. Lemma 3.1 is the key decomposition result. It is proved here and used in [ADLP] and [DW]. Section 4 introduces $\mathfrak{g}$-semistandard posets, $\mathfrak{g}$-semistandard lattices, and $\mathfrak{g}$-semistandard tableaux. Section 5 shows that the elements of these lattices match up with tableaux presented in Littelmann's [Lit]. This match-up yields our main results. Section 6 contains further remarks and problems.

2. Definitions and preliminary results. The reference for standard combinatorics material is [Sta2], and the reference for standard representation theory material is [Hum]. We use "R" (and "Q") as a generic name for most of the combinatorial structures defined in this section: "edge-colored directed graph," "vertex-colored directed graph," "ranked poset," and "splitting poset." The letter P is reserved for posets and vertex-colored posets that arise as posets of join irreducibles for distributive lattices. The letter L is reserved for distributive lattices and edge-colored distributive lattices. All posets are finite. We identify a poset with its Hasse diagram.

Let I be any set. An *edge-colored directed graph with edges colored by the set* I is a directed graph R with vertex set $\mathcal{V}(R)$ and directed-edge set $\mathcal{E}(R)$ together with a function $\mathbf{edgecolor}_R : \mathcal{E}(R) \longrightarrow I$ assigning to each edge of R a *color* from the set I. If an edge $\mathbf{s} \to \mathbf{t}$ in R is assigned $i \in I$, we write $\mathbf{s} \stackrel{i}{\to} \mathbf{t}$. See Figure 2.1. For $i \in I$, we let $\mathcal{E}_i(R)$ denote the set of edges in R of color i, so $\mathcal{E}_i(R) = \mathbf{edgecolor}_R^{-1}(i)$. If J is a subset of I, remove all edges from R whose colors are not in J; connected components of the resulting edge-colored directed graph are called J-*components* of R. For any $\mathbf{t}$ in R and any $J \subset I$, we let $\mathbf{comp}_J(\mathbf{t})$ denote the J-component of R containing $\mathbf{t}$. The *dual* R^* is the edge-colored directed graph whose vertex set $\mathcal{V}(R^*)$ is the set of symbols $\{\mathbf{t}^*\}_{\mathbf{t} \in \mathcal{V}(R)}$ together with colored edges $\mathcal{E}_i(R^*) := \{\mathbf{t}^* \stackrel{i}{\to} \mathbf{s}^* \,|\, \mathbf{s} \stackrel{i}{\to} \mathbf{t} \in \mathcal{E}_i(R)\}$ for each $i \in I$. Let Q be another edge-colored directed graph with edge colors from I. If R and Q have disjoint vertex sets, then the *disjoint sum* $R \oplus Q$ is the expected edge-colored directed graph. If $\mathcal{V}(Q) \subseteq \mathcal{V}(R)$ and $\mathcal{E}_i(Q) \subseteq \mathcal{E}_i(R)$ for each $i \in I$, then Q is an *edge-colored subgraph* of R. Let $R \times Q$ denote the expected edge-colored directed graph with vertex set $\mathcal{V}(R) \times \mathcal{V}(Q)$. The notion of isomorphism for edge-colored directed graphs is as expected. (See [ADLP] if any "expected" statement is unclear.) If R is an edge-colored directed graph with edges colored by the set I, and if $\sigma : I \longrightarrow I'$ is a mapping of sets, then we let R^σ be the edge-colored directed graph with edge color function $\mathbf{edgecolor}_{R^\sigma} := \sigma \circ \mathbf{edgecolor}_R$. We call R^σ a *recoloring* of R. Observe that $(R^*)^\sigma \cong (R^\sigma)^*$. We similarly define a *vertex-colored directed graph* with a function $\mathbf{vertexcolor}_R : \mathcal{V}(R) \longrightarrow I$ that assigns colors to the vertices of R. In this context, we speak of the *dual vertex-colored directed graph* R^*, the *disjoint sum* of two vertex-colored directed graphs with disjoint vertex sets, the *isomorphism* of vertex-colored directed graphs, *recoloring*, etc. For $\mathbf{s}$ and $\mathbf{t}$ in a poset R, there is a directed edge $\mathbf{s} \to \mathbf{t}$ in the Hasse diagram of R if and only if $\mathbf{t}$ *covers* $\mathbf{s}$. So terminology that applies to directed graphs (*connected*, *edge-colored*, *dual*, *vertex-colored*, etc.) will also apply to posets. The vertex $\mathbf{s}$ and the edge $\mathbf{s} \to \mathbf{t}$ are *below* $\mathbf{t}$, and the vertex $\mathbf{t}$ and the edge $\mathbf{s} \to \mathbf{t}$ are *above* $\mathbf{s}$. The vertex $\mathbf{s}$ is a *descendant* of $\mathbf{t}$, and $\mathbf{t}$ is an *ancestor* of $\mathbf{s}$. All edge-colored and vertex-colored directed graphs in this paper will turn out to be posets. See Figures 2.1 and 2.2.

For a directed graph R, a *rank function* is a surjective function $\rho : R \longrightarrow \{0, \ldots, l\}$ (where $l \geq 0$) with the property that if $\mathbf{s} \to \mathbf{t}$ in R, then $\rho(\mathbf{s}) + 1 = \rho(\mathbf{t})$. If such a

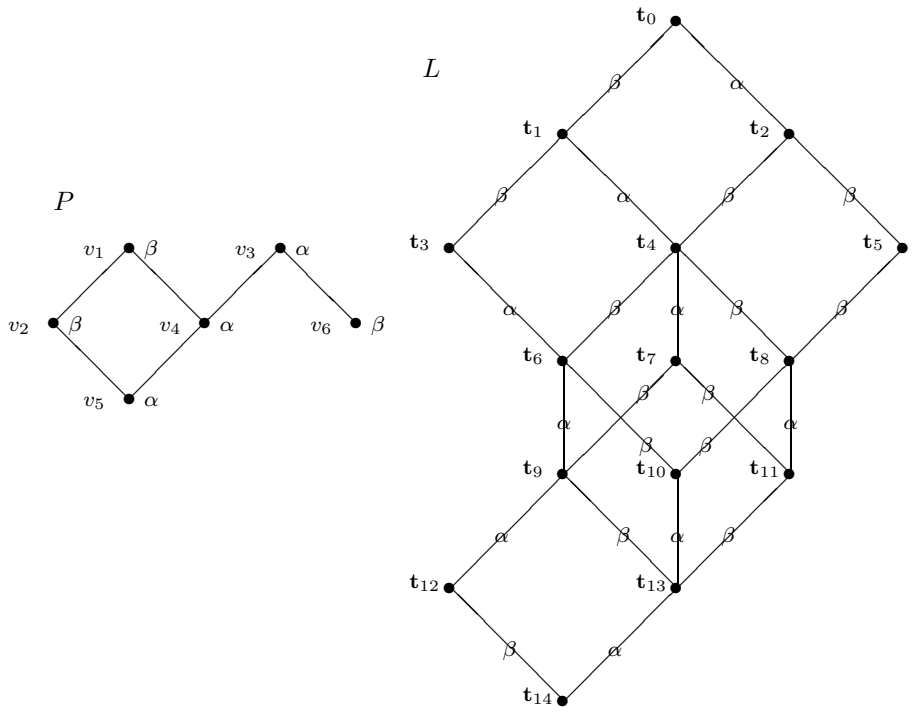


FIG. 2.1. *A vertex-colored poset P and an edge-colored lattice L.*

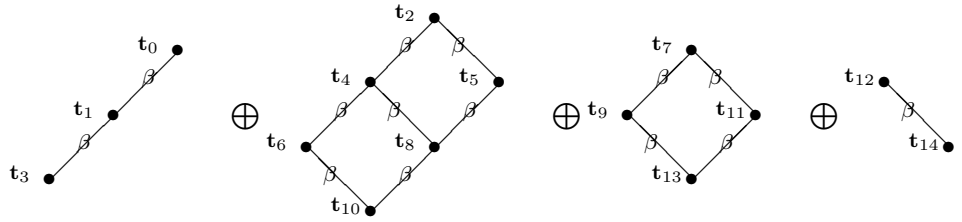


FIG. 2.2. *The disjoint sum of the β-components of the edge-colored lattice L from Figure* 2.1.

rank function exists, then R is the Hasse diagram for a poset—a *ranked* poset. We call l the *length* of R with respect to ρ, and the set $\rho^{-1}(i)$ is the *ith rank* of R. In an edge-colored ranked poset R, $\mathbf{comp}_i(\mathbf{t})$ will be a ranked poset for each $\mathbf{t} \in R$ and $i \in I$. We let $l_i(\mathbf{t})$ denote the length of $\mathbf{comp}_i(\mathbf{t})$, and we let $\rho_i(\mathbf{t})$ denote the rank of $\mathbf{t}$ within this component. We define the *depth* of $\mathbf{t}$ in its i-component to be $\delta_i(\mathbf{t}) := l_i(\mathbf{t}) - \rho_i(\mathbf{t})$. A ranked poset R with rank function ρ and length l is *rank symmetric* if $|\rho^{-1}(i)| = |\rho^{-1}(l-i)|$ for $0 \le i \le l$. It is *rank unimodal* if there is an m such that $|\rho^{-1}(0)| \le |\rho^{-1}(1)| \le \cdots \le |\rho^{-1}(m)| \ge |\rho^{-1}(m+1)| \ge \cdots \ge |\rho^{-1}(l)|$.

The distributive lattice of order ideals of a poset P, partially ordered by subset containment, will be denoted $J(P)$. See [Sta2]. A coloring of the vertices of the poset P gives a natural coloring of the edges of the distributive lattice $L = J(P)$ as follows: Given a function $\mathbf{vertexcolor}_P : \mathcal{V}(P) \longrightarrow I$, we assign a covering relation $\mathbf{s} \to \mathbf{t}$ in L the color i and write $\mathbf{s} \stackrel{i}{\to} \mathbf{t}$ if $\mathbf{t} \setminus \mathbf{s} = \{u\}$ and $\mathbf{vertexcolor}_P(u) = i$. So L becomes

an edge-colored distributive lattice with edges colored by the set I; we write $L = J_{color}(P)$. The edge-colored lattice $L_{G_2}(0,1)$ of Figure 4.3 is obtained from the vertex-colored poset $P_{G_2}(0,1)$ of Figure 4.2 in this way. Note that $J_{color}(P^*) \cong (J_{color}(P))^*$, $J_{color}(P^\sigma) \cong (J_{color}(P))^\sigma$ (recoloring), and $J_{color}(P \oplus Q) \cong J_{color}(P) \times J_{color}(Q)$. An edge-colored poset P has the *diamond coloring property* if whenever

is an edge-colored subgraph of the Hasse diagram for P, then $i = l$ and $j = k$. A necessary and sufficient condition for an edge-colored distributive lattice L to be isomorphic (as an edge-colored poset) to $J_{color}(P)$ for some vertex-colored poset P is for L to have the diamond coloring property. Then for $\mathbf{s} \in L$ and $i \in I$, one can see that $\mathbf{comp}_i(\mathbf{s})$ is the Hasse diagram for a distributive lattice. In particular, $\mathbf{comp}_i(\mathbf{s})$ is a distributive sublattice of L in the induced order, and a covering relation in $\mathbf{comp}_i(\mathbf{s})$ is also a covering relation in L.

Let $n \geq 1$. Let $\mathcal{D}$ be a Dynkin diagram with n nodes which are indexed by the elements of a set I such that $|I| = n$. The associated Cartan matrix is denoted $(\mathcal{D}_{i,j})_{i,j \in I}$. Throughout this paper $\mathfrak{g}$ will denote the complex semisimple Lie algebra of rank n with Chevalley generators $\{x_i, y_i, h_i\}_{i \in I}$ satisfying the Serre relations specified by the Cartan matrix for the Dynkin diagram at hand. Usually $I = \{1, \ldots, n\}$. In any Cartan matrix, $\mathcal{D}_{i,i} = 2$ for $i \in I$. Figure 2.3 presents the off-diagonal entries $\mathcal{D}_{i,j}$, $i \neq j$, for the rank two semisimple Dynkin diagrams $A_1 \oplus A_1$, A_2, C_2, and G_2. Two Dynkin diagrams $\mathcal{D}$ and $\mathcal{D}'$ are *isomorphic* if under some one-to-one correspondence $\sigma : I \longrightarrow I'$ we have $\mathcal{D}_{i,j} = \mathcal{D}'_{\sigma(i),\sigma(j)}$ and $\mathcal{D}_{j,i} = \mathcal{D}'_{\sigma(j),\sigma(i)}$. Let E denote the Euclidean space equipped with an inner product $\langle \cdot, \cdot \rangle$ which contains the root system Φ associated to $\mathcal{D}$. The set of simple roots is denoted $\{\alpha_i\}_{i \in I}$. For a root α, the *coroot* is $\alpha^\vee := \frac{2\alpha}{\langle \alpha, \alpha \rangle}$. The (i,j)-element $\mathcal{D}_{i,j}$ of the Cartan matrix is $\langle \alpha_i, \alpha_j^\vee \rangle$. The *fundamental weights* $\{\omega_1, \ldots, \omega_n\}$ form the basis for E dual to the simple coroots $\{\alpha_i^\vee\}_{i=1}^n$: $\langle \omega_j, \alpha_i^\vee \rangle = \delta_{i,j}$. The *lattice of weights* Λ is the set of all integral linear combinations of the fundamental weights. We coordinatize Λ to obtain a one-to-one correspondence with $\mathbb{Z}^n$ as follows: identify ω_i with the axis vector $(0, \ldots, 1, \ldots, 0)$, where "1" is in the ith position. For $i \in I$, $\alpha_i = \sum_{j \in I} \mathcal{D}_{i,j} \omega_j$. So the simple root α_i can be identified with the ith row vector of the Cartan matrix. The *Weyl group* W is generated by the *simple reflections* $s_i : E \to E$ for all $i \in I$: Here $s_i(v) = v - \langle v, \alpha_i^\vee \rangle \alpha_i$ for $v \in E$.

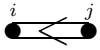
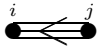
Subgraph				
$\mathcal{D}_{i,j}$, $\mathcal{D}_{j,i}$	0 , 0	−1 , −1	−1 , −2	−1 , −3

FIG. 2.3.

Vector spaces in this paper are complex and finite dimensional. If V is a $\mathfrak{g}$-module, then there is at least one basis $\mathcal{B} := \{v_\mathbf{s}\}_{\mathbf{s} \in R}$ (where R is an indexing set with $|R| = \dim V$) consisting of eigenvectors for the actions of the h_i's: for any $\mathbf{s}$ in R and $i \in I$, there exists an integer $k_i(\mathbf{s})$ such that $h_i.v_\mathbf{s} = k_i(\mathbf{s}) v_\mathbf{s}$. The *weight* of the basis vector $v_\mathbf{s}$ is the sum $wt(v_\mathbf{s}) := \sum_{i \in I} k_i(\mathbf{s}) \omega_i$. We say that $\mathcal{B}$ is a *weight basis* for V. If μ is a weight in Λ, then we let V_μ be the subspace of V spanned by all basis vectors $v_\mathbf{s} \in \mathcal{B}$ such that $wt(v_\mathbf{s}) = \mu$. The subspace V_μ is independent of the choice of weight basis $\mathcal{B}$. The finite dimensional irreducible $\mathfrak{g}$-modules are indexed by their

"highest weights" λ as these highest weights λ run through the *dominant* weights Λ^+ (the nonnegative linear combinations of the fundamental weights). The Lie algebra $\mathfrak{g}$ acts on the dual space V^* by the rule $(z.f)(v) = -f(z.v)$ for all $v \in V$, $f \in V^*$, and $z \in \mathfrak{g}$.

Let R be a ranked poset whose Hasse diagram edges are colored with colors taken from I, $|I| = n$. For $i \in I$, find the connected components of the subgraph with edges $\mathcal{E}_i(R)$. For $i \in I$ and $\mathbf{s}$ in R, set $m_i(\mathbf{s}) := \rho_i(\mathbf{s}) - \delta_i(\mathbf{s}) = 2\rho_i(\mathbf{s}) - l_i(\mathbf{s})$. Let $wt_R(\mathbf{s})$ be the n-tuple $(m_i(\mathbf{s}))_{i\in I}$. See Figure 4.4. Given a matrix $M = (M_{p,q})_{p,q\in I}$, then for fixed $i \in I$ let $M^{(i)}$ be the n-tuple $(M_{i,j})_{j\in I}$, the ith row vector for M. We say that R *satisfies the structure condition for* M if $wt_R(\mathbf{s}) + M^{(i)} = wt_R(\mathbf{t})$ whenever $\mathbf{s} \xrightarrow{i} \mathbf{t}$ for some $i \in I$; that is, for all $j \in I$ we have $m_j(\mathbf{s}) + M_{i,j} = m_j(\mathbf{t})$. Following [DLP], we say that R *satisfies the* $\mathfrak{g}$*-structure condition* if M is the Cartan matrix for the Dynkin diagram $\mathcal{D}$ associated to $\mathfrak{g}$. In this case view $wt_R : R \longrightarrow \Lambda$ as the function given by $wt_R(\mathbf{s}) = \sum_{j\in I} m_j(\mathbf{s})\omega_j$. Then R satisfies the $\mathfrak{g}$-structure condition if and only if for each simple root α_i we have $wt_R(\mathbf{s}) + \alpha_i = wt_R(\mathbf{t})$ whenever $\mathbf{s} \xrightarrow{i} \mathbf{t}$ in R. (In [Don1] the edges of R were said to "preserve weights.") This condition requires the color structure of R to be compatible with the structure of the set of weights for a representation of $\mathfrak{g}$. The largest edge-colored distributive lattice of Figure 4.4 satisfies the structure condition for the G_2 Cartan matrix (Figure 4.1) and therefore satisfies the G_2-structure condition.

The following obvious lemma is used when the Dynkin diagram has symmetry or when other numberings of the Dynkin diagram are convenient.

LEMMA 2.1. *Let $\mathcal{D}$ and $\mathcal{D}'$ be Dynkin diagrams with nodes indexed by I and I' such that $\mathcal{D}$ and $\mathcal{D}'$ are isomorphic under a one-to-one correspondence $\sigma : I \longrightarrow I'$. Let $\mathfrak{g}$ and $\mathfrak{g}'$ be the respective semisimple Lie algebras. Let R be a ranked poset with edges colored by the set I, and consider the recoloring R^σ. Then R satisfies the $\mathfrak{g}$-structure condition if and only if R^σ satisfies the $\mathfrak{g}'$-structure condition.*

Let w_0 be the longest element of the Weyl group W associated to $\mathfrak{g}$, as in Exercise 10.9 of [Hum]. When w_0 acts on Λ, then for each i it sends $\alpha_i \mapsto -\alpha_{\sigma_0(i)}$ and $\omega_i \mapsto -\omega_{\sigma_0(i)}$, where $\sigma_0 : I \longrightarrow I$ is some permutation of the node labels of the Dynkin diagram $\mathcal{D}$. Here σ_0 must be a symmetry of the Dynkin diagram, and since $w_0^2 = id$ in W it is the case that σ_0^2 is the identity permutation. For any weight $\mu = \sum a_i\omega_i$ we have $-w_0\mu = \sum a_i\omega_{\sigma_0(i)}$. For connected Dynkin diagrams, σ_0 is trivial except in the cases A_n $(n \geq 2)$, D_{2k+1} $(k \geq 2)$, and E_6; in these cases it is the only nontrivial Dynkin diagram automorphism. Given an edge-colored poset R with edges colored by the set I of indices for the Dynkin diagram $\mathcal{D}$, we let $R^\triangle$ be the edge-colored poset $(R^*)^{\sigma_0}$ and call $R^\triangle$ the σ_0*-recolored dual* of R. Observe that $(R^\triangle)^\triangle = R$. We allow "$\triangle$" to be applied to any vertex-colored poset Q whose vertex colors correspond to nodes of a Dynkin diagram.

The group ring $\mathbb{Z}[\Lambda]$ has vector space basis $\{e_\mu \,|\, \mu \in \Lambda\}$ and multiplication rule $e_{\mu+\nu} = e_\mu e_\nu$. The Weyl group W acts on $\mathbb{Z}[\Lambda]$ by the rule $\sigma.e_\mu := e_{\sigma\mu}$. The *character ring* $\mathbb{Z}[\Lambda]^W$ for $\mathfrak{g}$ is the ring of W-invariant elements of $\mathbb{Z}[\Lambda]$; elements of $\mathbb{Z}[\Lambda]^W$ are *characters for* $\mathfrak{g}$. If V is a representation of $\mathfrak{g}$, then the *Weyl character for* V is $\chi(V) := \sum_{\mu\in\Lambda} (\dim V_\mu)e_\mu \in \mathbb{Z}[\Lambda]^W$. If V is irreducible with highest weight λ, let $\chi_\lambda := \chi(V)$. We call χ_λ an *irreducible character*. Let $A_\mu := \sum_{\sigma\in W} \det(\sigma)e_{\sigma\mu}$. Let $\varrho := \omega_1 + \cdots + \omega_n$. It is well known that $A_\varrho = e_\varrho \Pi(1 - e_{-\alpha})$, product taken over the positive roots α. Weyl's character formula says that χ_λ is the unique element of $\mathbb{Z}[\Lambda]^W$ for which $A_\varrho\chi_\lambda = A_{\varrho+\lambda}$.

Let V be a representation of $\mathfrak{g}$. A *splitting system for* V (or for $\chi(V)$) is a

pair $(\mathcal{T}, weight)$, where $\mathcal{T}$ is a set and $weight : \mathcal{T} \longrightarrow \Lambda$ is a *weight function* such that $\chi(V) := \sum_{\mathbf{t}\in\mathcal{T}} e_{weight(\mathbf{t})}$. If R is a ranked poset with edges colored by the set $\{1, \ldots, n\}$, if R satisfies the structure condition for $\mathfrak{g}$, and if (R, wt_R) is a splitting system for V, then we say that R is a *splitting poset for* V (or for $\chi(V)$). This concept appears unnamed on page 266 of [Don1] and as "labeling poset" in Corollary 5.3 of [ADLP]. An edge-colored ranked poset R for which (R, wt_R) is a splitting system for an irreducible representation can fail to satisfy the structure condition for $\mathfrak{g}$. We use z_i to denote e_{ω_i}. If R is a splitting poset for V, then $\chi(V) = \sum_{\mathbf{t}\in R}(z_1, \ldots, z_n)^{wt_R(\mathbf{t})}$, where $(z_1, \ldots, z_n)^{wt_R(\mathbf{t})} := z_1^{m_1(\mathbf{t})} \cdots z_n^{m_n(\mathbf{t})}$. Here χ_λ is a Laurent polynomial in the indeterminates z_i with nonnegative integer coefficients. We denote this polynomial by $char_{\mathfrak{g}}(\lambda; z_1, \ldots, z_n)$.

LEMMA 2.2. *Let V be a representation for a semisimple Lie algebra $\mathfrak{g}$. Let $\mathfrak{g}'$ be a semisimple Lie algebra isomorphic to $\mathfrak{g}$ obtained from an isomorphism σ of Dynkin diagrams as in the statement of Lemma 2.1. Suppose R is a splitting poset for V. Then the edge-colored poset R^* is a splitting poset for the dual representation V^* of $\mathfrak{g}$, R^σ is a splitting poset for the $\mathfrak{g}'$-module V, and $R^\triangle$ is a splitting poset for the $\mathfrak{g}$-module V.*

Proof. The only assertion that does not immediately follow from the definitions and Lemma 2.1 is that $R^\triangle$ is a splitting poset *for the* $\mathfrak{g}$*-module* V. Write $V \cong V_1 \oplus \cdots \oplus V_k$, a decomposition of V into irreducible $\mathfrak{g}$-modules V_i such that V_i has highest weight μ_i. The dual $\mathfrak{g}$-module V^* has R^* as a supporting graph; V^* decomposes as $V_1^* \oplus \cdots \oplus V_k^*$, where each V_i^* is irreducible with highest weight $-w_0(\mu_i)$ (cf. Exercise 21.6 of [Hum]). Recolor R^* by applying the permutation σ_0 to obtain $R^\triangle$. Now view V^* as a new $\mathfrak{g}$-module U induced by the action $x_i.v := x_{\sigma_0(i)}.v$ and $y_i.v := y_{\sigma_0(i)}.v$ for each $i \in I$ and $v \in V^*$. It is apparent that $R^\triangle$ is a splitting poset for the $\mathfrak{g}$-module U. Let U_i be the (irreducible) $\mathfrak{g}$-submodule of U corresponding to V_i^*. One can see that the highest weight of U_i is now $-w_0(-w_0(\mu_i))$, which is just μ_i. Hence U is isomorphic to V. □

LEMMA 2.3. *Let V be an irreducible $\mathfrak{g}$-module. Then there is a connected splitting poset for V.*

Proof. By Lemmas 3.1.A, 3.1.F, and 3.2.A of [Don1], any supporting graph for V will do. □

This paragraph and Proposition 2.4 borrow from sections 5 and 6 of [Pr1]. If we set

$$x := 2\sum_{i=1}^{n}\left[\sum_{j=1}^{n}\frac{2\langle\omega_i,\omega_j\rangle}{\langle\alpha_j,\alpha_j\rangle}\right]x_i, \quad y := \sum y_i, \quad \text{and} \quad h := 2\sum_{i=1}^{n}\left[\sum_{j=1}^{n}\frac{2\langle\omega_i,\omega_j\rangle}{\langle\alpha_j,\alpha_j\rangle}\right]h_i,$$

then $\mathfrak{s} := \mathrm{span}\{x, y, h\}$ is a three-dimensional subalgebra of $\mathfrak{g}$ isomorphic to $\mathfrak{sl}(2, \mathbb{C})$. It is called a "principal three-dimensional subalgebra." Set $\varrho^\vee := \sum_{i=1}^{n}\frac{2\omega_i}{\langle\alpha_i,\alpha_i\rangle}$. Observe that $\langle\alpha_i, \varrho^\vee\rangle = 1$ for $1 \le i \le n$. Let V be a $\mathfrak{g}$-module. Let R be a splitting poset for V. Then there exists a weight basis for V which can be indexed by the elements of R, say, $\{v_\mathbf{t}\}_{\mathbf{t}\in R}$, so that the weight of the basis vector $v_\mathbf{t}$ is $wt_R(\mathbf{t})$. One can check that $h.v_\mathbf{t} = 2\langle wt_R(\mathbf{t}), \varrho^\vee\rangle v_\mathbf{t}$, so the set $\{2\langle wt_R(\mathbf{x}), \varrho^\vee\rangle\}_{\mathbf{x}\in R}$ consists of the integral weights for V regarded as an $\mathfrak{s}$-module. Choose an element $\mathbf{max}$ in R such that $2\langle wt_R(\mathbf{max}), \varrho^\vee\rangle$ is the largest in the set $\{2\langle wt_R(\mathbf{x}), \varrho^\vee\rangle\}_{\mathbf{x}\in R}$, and choose $\mathbf{min}$ such that $2\langle wt_R(\mathbf{min}), \varrho^\vee\rangle$ is the smallest. Symmetry of the integral weights for V under the action of $\mathfrak{s} \cong \mathfrak{sl}(2, \mathbb{C})$ implies that $2\langle wt_R(\mathbf{max}), \varrho^\vee\rangle = -2\langle wt_R(\mathbf{min}), \varrho^\vee\rangle$. Set $l := 2\langle wt_R(\mathbf{max}), \varrho^\vee\rangle$. Since R satisfies the $\mathfrak{g}$-structure condition, it follows that if $\mathbf{s} \xrightarrow{i} \mathbf{t}$ is an edge in R, then

$wt_R(\mathbf{s}) + \alpha_i = wt_R(\mathbf{t})$; therefore, $\langle wt_R(\mathbf{s}), \varrho^\vee\rangle + 1 = \langle wt_R(\mathbf{t}), \varrho^\vee\rangle$. Suppose for the moment that R is connected. Then the weights $\{2\langle wt_R(\mathbf{x}), \varrho^\vee\rangle\}_{\mathbf{x}\in R}$ all have the same parity. Consider the function $\rho : R \longrightarrow \mathbb{Z}$ given by $\rho(\mathbf{t}) := \frac{l}{2} + \langle wt_R(\mathbf{t}), \varrho^\vee\rangle$. Based on what we have seen so far, the range of ρ is the set of integers $\{0, \ldots, l\}$, and hence ρ is the rank function for R. Next consider the case that V is irreducible with highest weight λ. Then R need not be connected. However, since V has a connected splitting poset by Lemma 2.3, then the weights $\{2\langle wt_R(\mathbf{x}), \varrho^\vee\rangle\}_{\mathbf{x}\in R}$ all have the same parity. Thus the function $\rho : R \longrightarrow \mathbb{Z}$ given by $\rho(\mathbf{t}) := \frac{l}{2} + \langle wt_R(\mathbf{t}), \varrho^\vee\rangle$ will be a rank function for R with range $\{0, \ldots, l\}$. Call ρ the *natural rank function* for R. Since V is irreducible, we can see that **max** is the unique element of R with weight $wt_R(\mathbf{max}) = \lambda$. Hence $l = 2\langle\lambda, \varrho^\vee\rangle$. Next we define the *rank generating function* for R to be $RGF_{\mathfrak{g}}(\lambda, q) := \sum_{i=0}^{l} |\rho^{-1}(i)|q^i = \sum_{\mathbf{t}\in R} q^{\rho(\mathbf{t})}$. This is the usual rank generating function for the ranked poset R. We do not refer to R in the notation $RGF_{\mathfrak{g}}(\lambda, q)$ because we have $\left|\rho^{-1}(i)\right| = \left|\{\mathbf{t} \in R \,|\, \frac{l}{2} + \langle wt_R(\mathbf{t}), \varrho^\vee\rangle = i\}\right| = \sum_\mu \dim(V_\mu)$, where the latter sum is over all weights μ such that $\frac{l}{2} + \langle\mu, \varrho^\vee\rangle = i$. Thus if R' is another naturally ranked splitting poset for V, then corresponding ranks of R and R' have the same size. To obtain the rank generating function identity in the following result we use the "principal specialization" of Weyl's character formula from section 6 of [Pr1].

PROPOSITION 2.4. *Let V be an irreducible $\mathfrak{g}$-module with highest weight λ, and let R be a splitting poset for V with the natural rank function identified in the preceding paragraph. (If R is connected, then the natural rank function is the unique rank function.) Then R is rank symmetric and rank unimodal, and*

$$RGF_{\mathfrak{g}}(\lambda, q) = \frac{\Pi_{\alpha\in\Phi^+}(1 - q^{\langle\lambda+\varrho,\alpha^\vee\rangle})}{\Pi_{\alpha\in\Phi^+}(1 - q^{\langle\varrho,\alpha^\vee\rangle})}.$$

Proof. Choose a connected splitting poset R' for V; the natural rank function for R' is the unique rank function. Then by Proposition 3.5 of [Don1], it follows that R' is rank symmetric and rank unimodal. From the observation of the next-to-last sentence of the paragraph preceding the proposition, we conclude that the naturally ranked poset R is rank symmetric and rank unimodal. The principal specialization obtained from [Pr1, pp. 337–338] is for simple Lie algebras, but the same arguments are valid for semisimple Lie algebras. Apply this to obtain the rank generating function identity of the proposition statement. □

3. Grid posets and two-color grid posets. Here we introduce general grid posets and two-color grid posets with purely combinatorial definitions. From section 4 onward we will consider only the particular two-color grid posets called "$\mathfrak{g}$-semistandard" posets, whose structures are indexed by rank two Dynkin diagrams. Some (uncolored) grid posets are displayed in Figure 3.1; the poset P in Figure 2.1 is a two-color grid poset. In the general setting of this section, Lemma 3.1 and its related definitions provide for the decomposition of two-color grid posets into manageable pieces. Given $m \geq 1$, set $[m] := \{1, 2, \ldots, m\}$.

Given a finite poset $(P, \leq_P)$, a *chain function for P* is a function $\mathbf{chain} : P \longrightarrow [m]$ for some positive integer m such that (1) $\mathbf{chain}^{-1}(i)$ is a (possibly empty) chain in P for $1 \leq i \leq m$, and (2) given any cover $u \rightarrow v$ in P, it is the case that either $\mathbf{chain}(u) = \mathbf{chain}(v)$ or $\mathbf{chain}(u) = \mathbf{chain}(v) + 1$. A *grid poset* is a finite poset $(P, \leq_P)$ together with a chain function $\mathbf{chain} : P \longrightarrow [m]$ for some $m \geq 1$. Depending on context, the notation P can refer to the grid poset $(P, \leq_P, \mathbf{chain} : P \longrightarrow [m])$ or the underlying poset $(P, \leq_P)$. The conditions on **chain** imply that an element in a

grid poset covers no more than two elements and is covered by no more than two elements.[1] Observe that if i is the smallest (respectively, largest) integer such that $\mathbf{chain}^{-1}(i)$ is nonempty and if u is the maximal (respectively, minimal) element of $\mathbf{chain}^{-1}(i)$, then u is a maximal (respectively, minimal) element of the poset P. A grid poset P is *connected* if and only if the Hasse diagram for the poset P is connected. For $1 \leq i \leq m$ we set $\mathcal{C}_i := \mathbf{chain}^{-1}(i)$. When we depict grid posets, the chains $\mathcal{C}_i$ will be directed from SW to NE. See Figure 3.1.

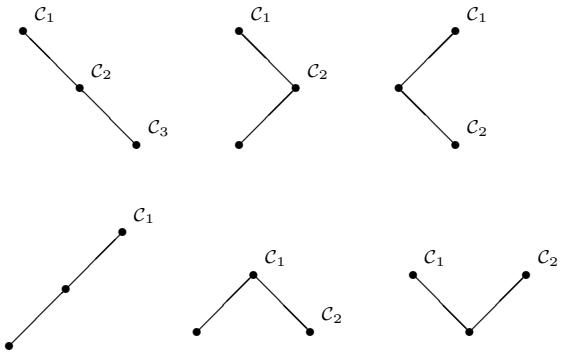


FIG. 3.1. *The six nonisomorphic connected grid posets with three elements.*

Let $(P, \leq_P, \mathbf{chain} : P \longrightarrow [m])$ be a grid poset. The *dual grid poset* P^* is the dual poset P^* together with the chain function $\mathbf{chain}^* : P^* \longrightarrow [m]$ given by $\mathbf{chain}^*(u^*) = m + 1 - \mathbf{chain}(u)$ for all $u \in P$. For $i = 1, 2$, let P_i be a grid poset with chain function $\mathbf{chain}_i : P_i \longrightarrow [m_i]$ for some $m_i \geq 1$. A one-to-one correspondence $\phi : P_1 \longrightarrow P_2$ is an *isomorphism of grid posets* if we have $u \to v$ in P_1 with $\mathbf{chain}_1(u) = \mathbf{chain}_1(v)$ (respectively, $\mathbf{chain}_1(u) = \mathbf{chain}_1(v) + 1$) if and only if $\phi(u) \to \phi(v)$ in P_2 with $\mathbf{chain}_2(\phi(u)) = \mathbf{chain}_2(\phi(v))$ (respectively, $\mathbf{chain}_2(\phi(u)) = \mathbf{chain}_2(\phi(v)) + 1$). Figure 3.1 depicts each of the isomorphism classes of connected grid posets with three elements apiece. Given a nonempty grid poset $(P, \leq_P, \mathbf{chain} : P \longrightarrow [m])$, there exists some $m' \geq 1$ and a surjective chain function $\mathbf{chain}' : P \longrightarrow [m']$ such that the grid poset P is isomorphic to $(P, \leq_P, \mathbf{chain}' : P \longrightarrow [m'])$. If P is connected, then this surjective chain function $\mathbf{chain}'$ is unique. We say that Q is a *grid subposet* of a given grid poset P if (1) Q is a subposet of P in the induced order, and (2) whenever $u \to v$ is a covering relation in Q then it is also a covering relation in P. In this case, we regard Q with the chain function $\mathbf{chain}|_Q$ to be a grid poset on its own.

For a grid poset $(P, \leq_P, \mathbf{chain} : P \longrightarrow [m])$, let $\mathcal{T}_P$ be the totally ordered set whose elements are the elements of P and whose ordering is given by the following rule: for distinct u and v in P write $u <_{\mathcal{T}_P} v$ if and only if (1) $\mathbf{chain}(u) < \mathbf{chain}(v)$ or (2) $\mathbf{chain}(u) = \mathbf{chain}(v)$ with $v <_P u$. Let $l := |P|$. Number the vertices of P $v_1, v_2, \ldots, v_l$ so that $v_p <_{\mathcal{T}_P} v_q$ whenever $1 \leq p < q \leq l$. Let $L := J(P)$ be the distributive lattice of order ideals of P. We simultaneously think of order ideals of P

[1] Motivation for terminology: For $m, n \geq 1$, let $\mathcal{G}$ be the directed graph with $\mathcal{V}(\mathcal{G}) = \{(p, q) \in \mathbb{Z} \times \mathbb{Z} | 1 \leq p \leq n, 1 \leq q \leq m\}$ and with $\mathcal{E}(\mathcal{G}) = \{(p, q) \to (r, s)$ if $(r, s) - (p, q) = (1, 0)$ or $(0, 1)\}$. We refer to $\mathcal{G}$ as a "directed grid graph." Here $\mathcal{G}$ is the Hasse diagram for a poset obtained by rotating the plane counterclockwise through an angle of 45° so that the vertex $(1, 1)$ of $\mathcal{G}$ is the minimal element. A grid poset $(P, \leq_P, \mathbf{chain} : P \longrightarrow [m])$ can be obtained as a subgraph of a directed grid graph for an appropriately large n by removing some vertices and some "NW" edges.

as subsets of P and as elements of L.

A *two-color function* for a grid poset $(P, \leq_P, \mathbf{chain} : P \longrightarrow [m])$ is a function $\mathbf{color} : P \longrightarrow \Delta$ such that (1) $|\Delta| = 2$, (2) $\mathbf{color}(u) = \mathbf{color}(v)$ if $\mathbf{chain}(u) = \mathbf{chain}(v)$, and (3) if u and v are in the same connected component of P with $\mathbf{chain}(u) = \mathbf{chain}(v) + 1$, then $\mathbf{color}(u) \neq \mathbf{color}(v)$. A *two-color grid poset* is a grid poset $(P, \leq_P, \mathbf{chain} : P \longrightarrow [m])$ together with a two-color function $\mathbf{color} : P \longrightarrow \Delta$. In some contexts we will use the notation P to refer to the two-color grid poset $(P, \leq_P, \mathbf{chain} : P \longrightarrow [m], \mathbf{color} : P \longrightarrow \Delta)$. Two-color grid posets are vertex-colored posets. To a two-color grid poset P we associate the edge-colored distributive lattice $L := J_{color}(P)$, as in section 2. The number of nonempty chains $\mathcal{C}_i$ in P of color $\gamma \in \Delta$ gives an upper bound for the number of ancestors (respectively, descendants) an element in L can have along edges of color γ. One can also see that any color γ component of L is poset-isomorphic to a product of chains. The *dual two-color grid poset* P^* is the dual grid poset P^* together with the two-color function $\mathbf{color}^* : P^* \longrightarrow \Delta$ given by $\mathbf{color}^*(u^*) = \mathbf{color}(u)$ for all $u \in P$. If Q is a grid subposet of the two-color grid poset P, then Q is a two-color grid poset with chain function $\mathbf{chain}|_Q$ and two-color function $\mathbf{color}|_Q$. In this case we call Q a *two-color grid subposet of* P. Two two-color grid posets $(P_i, \leq_{P_i}, \mathbf{chain}_i : P_i \longrightarrow [m_i], \mathbf{color}_i : P_i \longrightarrow \Delta)$ for $i = 1, 2$ are *isomorphic* if there is an isomorphism $\phi : P_1 \longrightarrow P_2$ of grid posets such that $\mathbf{color}_2(\phi(u)) = \mathbf{color}_1(u)$ for all u in P_1. We will often take $\Delta := \{\alpha, \beta\}$. When we *switch* (or *reverse*) the vertex colors of P we replace the color function $\mathbf{color} : P \longrightarrow \{\alpha, \beta\}$ with the color function $\mathbf{color}' : P \longrightarrow \{\alpha, \beta\}$ given by $\mathbf{color}'(v) = \alpha$ if $\mathbf{color}(v) = \beta$, and $\mathbf{color}'(v) = \beta$ if $\mathbf{color}(v) = \alpha$. Similarly, one can *switch* (or *reverse*) the edge colors of L. In Figures 3.2 and 3.3 we depict eight two-color grid posets; the numbering of the vertices for each poset P follows the total ordering $\mathcal{T}_P$. The vertex-colored poset P of Figure 2.1 is a two-color grid poset. The lattice L in that figure is $J_{color}(P)$.

In this paper the following definition is needed only for a comment in section 4 and for preview statements of Theorems 6.1 and 6.2. (It is also needed in [ADLP].) We say a two-color grid poset P has the *max property* if P is isomorphic to a two-color grid poset $(Q, \leq_Q, \mathbf{chain} : Q \longrightarrow [m], \mathbf{color} : Q \longrightarrow \Delta)$ with a surjective chain function such that (1) if u is any maximal element in the poset Q, then $\mathbf{chain}(u) \leq 2$, and (2) if $v \neq u$ is another maximal element in Q, then $\mathbf{color}(u) \neq \mathbf{color}(v)$. Note that the dual two-color grid poset P^* might fail to have the max property. The two-color grid posets of Figures 2.1, 3.2, and 3.3 have the max property.

Let P be a grid poset with chain function $\mathbf{chain} : P \longrightarrow [m]$. Suppose P_1 is a nonempty order ideal such that $P_1 \neq P$. Regard P_1 and $P_2 := P \setminus P_1$ to be subposets of the poset P in the induced order. Suppose that whenever u is a maximal (respectively, minimal) element of P_1 and v is a maximal (respectively, minimal) element of P_2, then $\mathbf{chain}(u) \leq \mathbf{chain}(v)$. Then we say that P *decomposes into* $P_1 \triangleleft P_2$, and we write $P = P_1 \triangleleft P_2$. If no such order ideal P_1 exists, then we say the grid poset P is *indecomposable*. See Figure 6.2. Note that if $P = P_1 \triangleleft P_2$ and $u < v$ in P with $u \in P_2$, then $v \in P_2$. Moreover, if $u \rightarrow v$ in P with $u \in P_1$ and $v \in P_2$, then $\mathbf{chain}(u) = \mathbf{chain}(v)$. Also, if $u \rightarrow v$ is a covering relation in the poset P_i for $i \in \{1, 2\}$, note that $u \rightarrow v$ is also a covering relation in P. Hence each P_i is a grid subposet of P. If P is a grid poset that decomposes into $P_1 \triangleleft Q$, and if Q decomposes into $P_2 \triangleleft P_3$, then $P = P_1 \triangleleft (P_2 \triangleleft P_3)$. But now observe that $P = (P_1 \triangleleft P_2) \triangleleft P_3$. So we may write $P = P_1 \triangleleft P_2 \triangleleft P_3$. In general, if $P = P_1 \triangleleft P_2 \triangleleft \cdots \triangleleft P_k$, then each P_i with chain function $\mathbf{chain}|_{P_i}$ is a grid subposet of P. Also, an order ideal $\mathbf{s}$ of P

may be expressed as the disjoint union $(\mathbf{s} \cap P_1) \cup (\mathbf{s} \cap P_2) \cup \cdots \cup (\mathbf{s} \cap P_k)$, where each $\mathbf{s} \cap P_i$ is an order ideal in P_i. If in addition $P = P_1 \lhd P_2 \lhd \cdots \lhd P_k$ is a two-color grid poset with two-color function **color**, then each P_i with chain function $\mathbf{chain}|_{P_i}$ and two-color function $\mathbf{color}|_{P_i}$ is a two-color grid subposet of P. Here $P_1 \lhd P_2 \lhd \cdots \lhd P_k$ is a decomposition of P into two-color grid posets.

Consider a two-color grid poset $(P, \leq_P, \mathbf{chain} : P \longrightarrow [m], \mathbf{color} : P \longrightarrow \{\alpha, \beta\})$ with edge-colored distributive lattice $L = J_{color}(P)$. For each $\mathbf{s}$ in L we can view the quantity $wt_L(\mathbf{s})$ as the pair $(2\rho_\alpha(\mathbf{s}) - l_\alpha(\mathbf{s}),\, 2\rho_\beta(\mathbf{s}) - l_\beta(\mathbf{s}))$ in $\mathbb{Z} \times \mathbb{Z}$. The mapping $wt_L : L \longrightarrow \mathbb{Z} \times \mathbb{Z}$ is the *lattice weight function* for L. If $P = P_1 \lhd P_2 \lhd \cdots \lhd P_k$, then for each i we let $L_i := J_{color}(P_i)$ be the edge-colored lattice for the two-color grid subposet P_i of P. Then wt_{L_i} denotes the lattice weight function for L_i. We let $\rho_\alpha^{(i)}$ and $l_\alpha^{(i)}$ (respectively, $\rho_\beta^{(i)}$ and $l_\beta^{(i)}$) denote the rank and length functions for color α (respectively, color β) for L_i.

LEMMA 3.1. *Let $(P, \leq_P, \mathbf{chain} : P \longrightarrow [m], \mathbf{color} : P \longrightarrow \{\alpha, \beta\})$ be a two-color grid poset, and suppose P decomposes into $P = P_1 \lhd P_2 \lhd \cdots \lhd P_k$. Keep the notation of the preceding paragraph.*

(1) *Let $\gamma \in \Delta = \{\alpha, \beta\}$, and let $\mathbf{s}$ be an element of $L = J_{color}(P)$. Then*

$$\rho_\gamma(\mathbf{s}) = \sum_{i=1}^{k} \rho_\gamma^{(i)}(\mathbf{s} \cap P_i), \quad l_\gamma(\mathbf{s}) = \sum_{i=1}^{k} l_\gamma^{(i)}(\mathbf{s} \cap P_i), \quad \textit{and} \quad wt_L(\mathbf{s}) = \sum_{i=1}^{k} wt_{L_i}(\mathbf{s} \cap P_i).$$

(2) *Consequently, if there is a 2×2 matrix $M = (M_{\iota,\kappa})_{(\iota,\kappa) \in \Delta \times \Delta}$ such that each edge-colored distributive lattice $L_i = J_{color}(P_i)$ satisfies the structure condition for M, then L satisfies the structure condition for M as well.*

Proof. First we show how (2) follows from (1). Given an edge $\mathbf{s} \xrightarrow{\gamma} \mathbf{t}$ in L, then it is the case that for some j with $1 \leq j \leq k$ we have $\mathbf{s} \cap P_j \xrightarrow{\gamma} \mathbf{t} \cap P_j$ in L_j, while for $1 \leq i \leq k$ with $i \neq j$ we have $\mathbf{s} \cap P_i = \mathbf{t} \cap P_i$. Since L_j satisfies the structure condition, we see that $wt_{L_j}(\mathbf{s} \cap P_j) + M^{(\gamma)} = wt_{L_j}(\mathbf{t} \cap P_j)$. For $i \neq j$ we have $wt_{L_i}(\mathbf{s} \cap P_i) = wt_{L_i}(\mathbf{t} \cap P_i)$. By (1) it follows that $wt_L(\mathbf{s}) + M^{(\gamma)} = wt_L(\mathbf{t})$.

The results in (1) for general k follow by induction once we prove the results for $k = 2$. So let $k = 2$, $\mathbf{s} \in L$, and $\gamma \in \{\alpha, \beta\}$. It suffices to show that $\rho_\gamma(\mathbf{s}) = \rho_\gamma^{(1)}(\mathbf{s} \cap P_1) + \rho_\gamma^{(2)}(\mathbf{s} \cap P_2)$ and $l_\gamma(\mathbf{s}) = l_\gamma^{(1)}(\mathbf{s} \cap P_1) + l_\gamma^{(2)}(\mathbf{s} \cap P_2)$. Let $\mathbf{r}_0, \mathbf{r}_1, \ldots$ be the sequence with $\mathbf{r}_0 := \mathbf{s}$ and $\mathbf{r}_{j+1} := \mathbf{r}_j \setminus \{v_{i_{j+1}}\}$, where $j \geq 0$ and $v_{i_{j+1}}$ is the smallest vertex in $\mathcal{T}_P$ of color γ that can be removed from $\mathbf{r}_j$ so that $\mathbf{r}_{j+1}$ is an order ideal of P. Let $\mathbf{r}_q$ be the terminal element of the sequence. Observe that $i_1 < i_2 < \cdots < i_q$. We have $\mathbf{r}_q \xrightarrow{\gamma} \mathbf{r}_{q-1} \xrightarrow{\gamma} \cdots \xrightarrow{\gamma} \mathbf{r}_1 \xrightarrow{\gamma} \mathbf{r}_0 = \mathbf{s}$. Similarly define a sequence $\mathbf{u}_0, \mathbf{u}_1, \ldots$, where $\mathbf{u}_0 := \mathbf{s}$ and $\mathbf{u}_{s+1} := \mathbf{u}_s \cup \{v_{r_{s+1}}\}$, where $s \geq 0$ and $v_{r_{s+1}}$ is the largest element in $\mathcal{T}_P$ of color γ not in $\mathbf{u}_s$ that can be added to $\mathbf{u}_s$ so that $\mathbf{u}_{s+1}$ is an order ideal of P. Let $\mathbf{u}_p$ be the terminal element of the sequence. Observe that $r_1 > r_2 > \cdots > r_p$. We have $\mathbf{s} = \mathbf{u}_0 \xrightarrow{\gamma} \cdots \xrightarrow{\gamma} \mathbf{u}_{p-1} \xrightarrow{\gamma} \mathbf{u}_p$. Since $\mathbf{comp}_\gamma(\mathbf{s})$ is the Hasse diagram for a distributive lattice, and since $\mathbf{r}_q$ and $\mathbf{u}_p$ are, respectively, a minimal and a maximal element in $\mathbf{comp}_\gamma(\mathbf{s})$, then it follows that $\mathbf{r}_q$ and $\mathbf{u}_p$ are, respectively, the unique minimal and the unique maximal elements of $\mathbf{comp}_\gamma(\mathbf{s})$. Then $\rho_\gamma(\mathbf{s}) = q$ and $l_\gamma(\mathbf{s}) = p + q$.

Reorganize the sequence $(v_{i_1}, \ldots, v_{i_q})$ as follows: write $(v_{k_1}, \ldots, v_{k_{q'}}, v_{k_{q'+1}}, \ldots, v_{k_q})$, where the vertices $v_{k_1}, \ldots, v_{k_{q'}}$ are all in P_2 with $k_1 < \cdots < k_{q'}$, and the vertices $v_{k_{q'+1}}, \ldots, v_{k_q}$ are all in P_1 with $k_{q'+1} < \cdots < k_q$. Set $\mathbf{r}'_0 := \mathbf{s}$ and for $j \geq 0$ set $\mathbf{r}'_{j+1} := \mathbf{r}'_j \setminus \{v_{k_{j+1}}\}$. We claim that each $\mathbf{r}'_{j+1}$ is an order ideal of P, and if $0 \leq j < q'$ (respectively, $q' \leq j < q$), then $v_{k_{j+1}}$ is the smallest element in $\mathcal{T}_P$ of color γ that is also in P_2 (respectively, P_1) that can be removed from $\mathbf{r}'_j$ so that

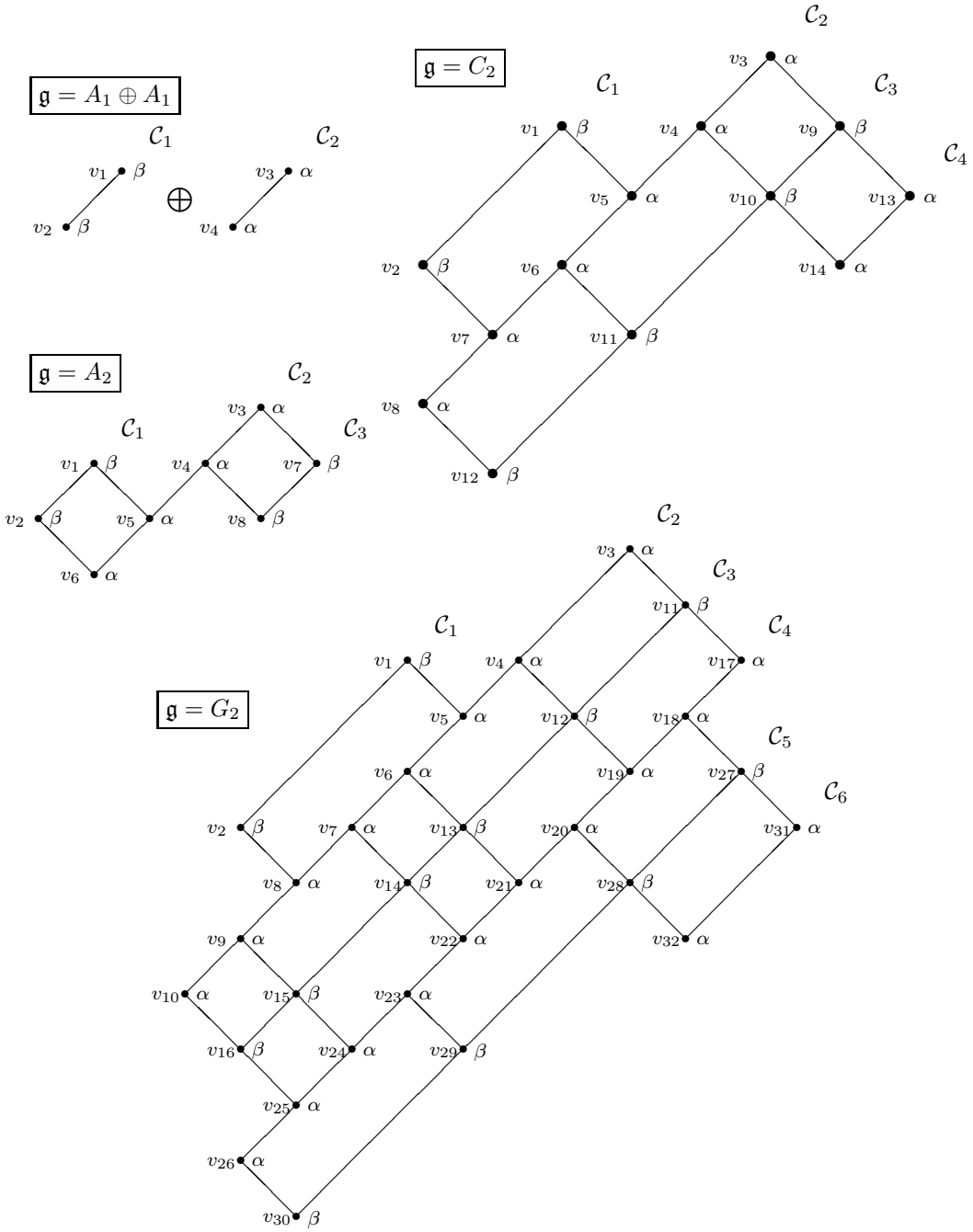


FIG. 3.2. *Depicted above are four two-color grid posets each possessing the max property. (Each is the* $\mathfrak{g}$*-semistandard poset* $P_{\mathfrak{g}}^{\beta\alpha}(2,2)$ *of section 4 for the indicated rank two semisimple Lie algebra* $\mathfrak{g}$*.)*

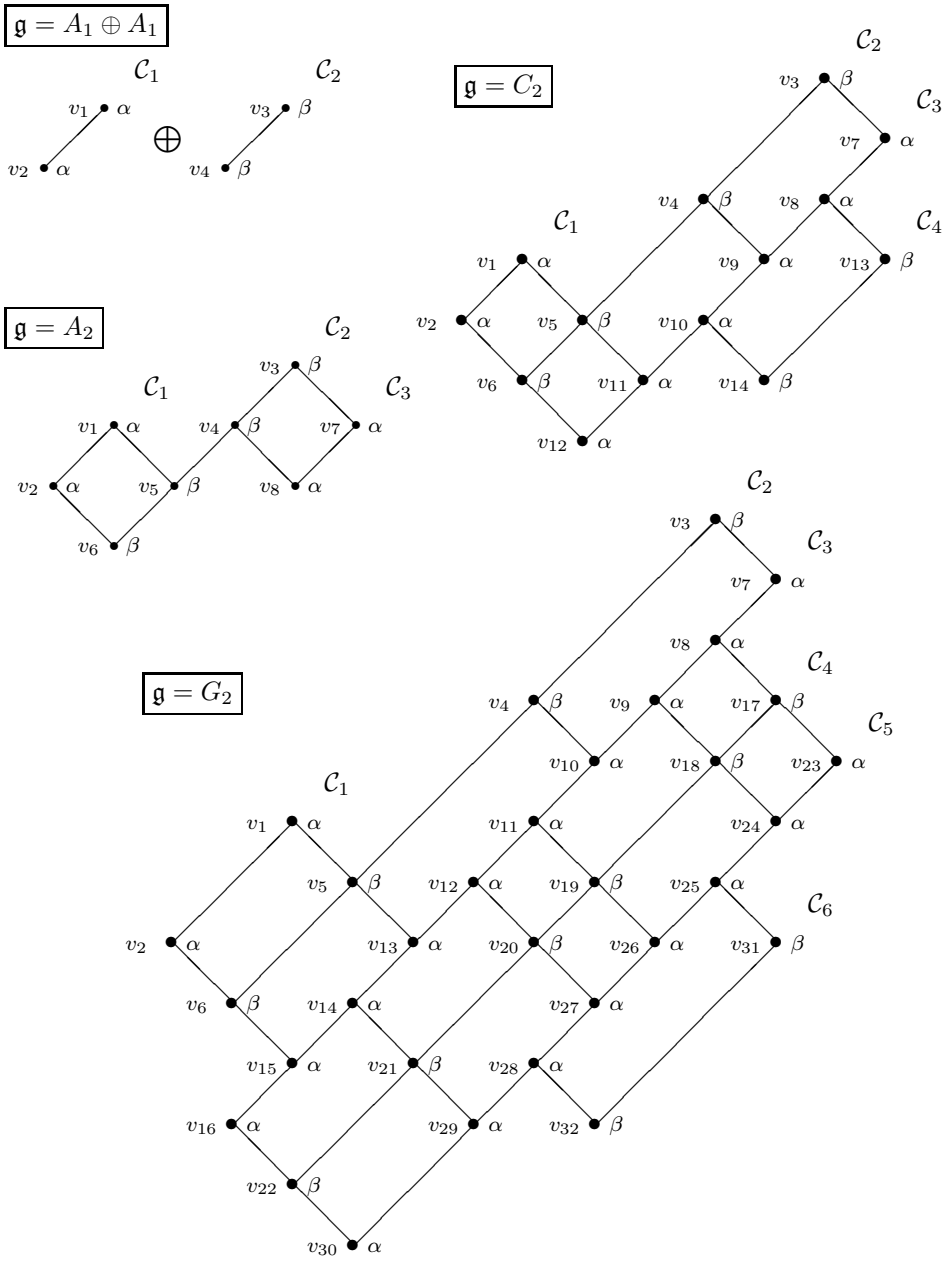


FIG. 3.3. *Depicted above are four two-color grid posets each possessing the max property. (Each is the* $\mathfrak{g}$*-semistandard poset* $P_{\mathfrak{g}}^{\alpha\beta}(2,2)$ *of section* 4 *for the indicated rank two semisimple Lie algebra* $\mathfrak{g}$*.)*

$\mathbf{r}'_{j+1}$ is an order ideal of P. (If so, we have a path $\mathbf{r}_q = \mathbf{r}'_q \xrightarrow{\gamma} \mathbf{r}'_{q-1} \xrightarrow{\gamma} \cdots \xrightarrow{\gamma} \mathbf{r}'_1 \xrightarrow{\gamma} \mathbf{r}'_0 = \mathbf{r}_0 = \mathbf{s}$ in L.) Proceed by induction on j. The statement follows if we can show that $v_{k_{j+1}}$ is maximal in $\mathbf{r}'_j$. First suppose $0 \leq j < q'$. If $v_{k_{j+1}}$ is not maximal in $\mathbf{r}'_j$, then $v_{k_{j+1}} < v$ for some other maximal element v in $\mathbf{r}'_j$. It must be the case that v is one of $(v_{k_{j+2}}, \ldots, v_{k_{q'}}, v_{k_{q'+1}}, \ldots, v_{k_q})$; otherwise one could not descend from $\mathbf{r}'_j$ to $\mathbf{r}_q$ in $L = J_{color}(P)$ along edges corresponding to vertices from $(v_{k_{j+2}}, \ldots, v_{k_{q'}}, v_{k_{q'+1}}, \ldots, v_{k_q})$. It cannot be the case that v is one of $(v_{k_{j+2}}, \ldots, v_{k_{q'}})$; otherwise $k_1 < \cdots < k_{j+1} < k_{j+2} < \cdots < k_{q'}$ implies that v is larger than $v_{k_{j+1}}$ in the total order $\mathcal{T}_P$, violating the fact that $v_{k_{j+1}} < v$ in P. And it cannot be the case that v is one of $(v_{k_{q'+1}}, \ldots, v_{k_q})$ since these are elements of P_1 and $v_{k_{j+1}}$ is in P_2. So for $0 \leq j < q'$, the vertex $v_{k_{j+1}}$ is maximal in $\mathbf{r}'_j$. Second, suppose that $q' \leq j < q$. If $v_{k_{j+1}}$ is not maximal in $\mathbf{r}'_j$, then by reasoning similar to the preceding case we have $v_{k_{j+1}} < v$ for some element v from $(v_{k_{j+2}}, \ldots, v_{k_q})$. But this violates the fact that $v_{k_{j+1}}$ precedes v in the total order $\mathcal{T}_P$ since $k_{q'} < \cdots < k_{j+1} < k_{j+2} < \cdots < k_q$. So for $q' \leq j < q$, the vertex $v_{k_{j+1}}$ is maximal in $\mathbf{r}'_j$. This concludes our induction on j.

Let $\mathbf{r}^{(1)}$ be the unique minimal element in the γ-component $\mathbf{comp}^{(1)}_\gamma(\mathbf{s} \cap P_1)$ of $\mathbf{s} \cap P_1$ in the edge-colored distributive lattice $L_1 = J_{color}(P_1)$. We claim that $\mathbf{r}^{(1)} = \mathbf{x}$, where $\mathbf{x} := (\mathbf{s} \cap P_1) \setminus \{v_{k_{q'+1}}, \ldots, v_{k_q}\}$. Now $\mathbf{x}$ is an order ideal of P_1 since $\mathbf{x} = \mathbf{r}'_q \cap P_1 = \mathbf{r}_q \cap P_1$. Also, $\mathbf{x} \in \mathbf{comp}^{(1)}_\gamma(\mathbf{s} \cap P_1)$ since $\mathbf{s} \cap P_1 = \mathbf{r}'_{q'} \cap P_1$ and the path $\mathbf{x} \xrightarrow{\gamma} (\mathbf{r}'_{q-1} \cap P_1) \xrightarrow{\gamma} \cdots \xrightarrow{\gamma} (\mathbf{r}'_{q'+1} \cap P_1) \xrightarrow{\gamma} (\mathbf{r}'_{q'} \cap P_1)$ stays in $\mathbf{comp}^{(1)}_\gamma(\mathbf{s} \cap P_1)$. If $\mathbf{r}^{(1)} \neq \mathbf{x}$, then $\mathbf{r}^{(1)} < \mathbf{x}$. In this case let $u \in \mathbf{x}$ be any color γ vertex such that $\mathbf{x} \setminus \{u\}$ is an order ideal of P_1. Let $\mathcal{C}_i$ be the chain in P that contains u. Note that u is not maximal in $\mathbf{r}_q \subseteq P$, and hence $u \to u'$ is a covering relation in P for some u' in $\mathbf{r}_q$. We refer to the following as observation $(*)$: If w is any element of P such that $u \to w$ and $w \in \mathbf{r}_q$, then $w \in P_2$. (Otherwise $w \in P_1$, so that $w \in \mathbf{x}$, and then $\mathbf{x} \setminus \{u\}$ cannot be an order ideal of P_1.) In particular, $u' \in P_2$. We claim that $u' \notin \{v_{k_1}, \ldots, v_{k_{q'}}\}$. Indeed, if $u' \in \{v_{k_1}, \ldots, v_{k_{q'}}\}$, then since $u \notin \{v_{k_{q'+1}}, \ldots, v_{k_q}\}$, it must be the case that $u \to u''$ for some $u'' \in \mathbf{r}_q$ in $\mathcal{C}_{i-1}$. By observation $(*)$, the element u'' is in P_2. But a covering relation in P between elements of P_1 and elements of P_2 can occur only along the chains $\mathcal{C}_1, \ldots, \mathcal{C}_m$. Therefore, $u'' \in \mathcal{C}_i$, which contradicts the fact that $u'' \in \mathcal{C}_{i-1}$. So it must be the case that $u' \notin \{v_{k_1}, \ldots, v_{k_{q'}}\}$. It follows that $u' < u''$ for some $u'' \in \mathbf{r}_q$ in $\mathcal{C}_{i-1}$. Let v be a maximal element in P such that $u'' \leq v$. Note that $v \in P_2$ since $u'' \in P_2$. Moreover, $v \in \mathcal{C}_j$ with $j \leq i-1$. Next suppose $u \to z$ for some $z \in P_1$. Since $z \in P_1$, then $z \neq u'$. Therefore, $z \in \mathcal{C}_{i-1}$. Therefore, $z \leq u''$. But since u'' is in the order ideal $\mathbf{r}_q$, it follows that $z \in \mathbf{r}_q$. But by observation $(*)$, it now follows that $z \in P_2$. This contradicts our hypothesis that $z \in P_1$. In particular, u must be a maximal element in P_1. So $u \in \mathcal{C}_i$ is a maximal element in P_1 and $v \in \mathcal{C}_j$ is a maximal element in P_2, and $j < i$. This violates the fact that P decomposes into $P_1 \triangleleft P_2$. So $\mathbf{r}^{(1)} = \mathbf{x}$, and hence $\rho^{(1)}_\gamma(\mathbf{s} \cap P_1) = q - q'$.

Let $\mathbf{r}^{(2)}$ be the unique minimal element in the γ-component $\mathbf{comp}^{(2)}_\gamma(\mathbf{s} \cap P_2)$ of $\mathbf{s} \cap P_2$ in the edge-colored distributive lattice $L_2 = J_{color}(P_2)$. We claim that $\mathbf{r}^{(2)} = \mathbf{y}$, where $\mathbf{y} := (\mathbf{s} \cap P_2) \setminus \{v_{k_1}, \ldots, v_{k_{q'}}\}$. Now $\mathbf{y}$ is an order ideal of P_2 since $\mathbf{y} = \mathbf{r}'_{q'} \cap P_2 = \mathbf{r}_q \cap P_2$. Also, $\mathbf{y} \in \mathbf{comp}^{(2)}_\gamma(\mathbf{s} \cap P_2)$ since the path $\mathbf{y} \xrightarrow{\gamma} (\mathbf{r}'_{q'-1} \cap P_2) \xrightarrow{\gamma} \cdots \xrightarrow{\gamma} (\mathbf{r}'_1 \cap P_2) \xrightarrow{\gamma} (\mathbf{r}'_0 \cap P_2)$ stays in $\mathbf{comp}^{(2)}_\gamma(\mathbf{s} \cap P_2)$. If $\mathbf{r}^{(2)} \neq \mathbf{y}$, then $\mathbf{r}^{(2)} < \mathbf{y}$. In this case let $u \in \mathbf{y}$ be any color γ vertex such that $\mathbf{y} \setminus \{u\}$ is an order ideal of P_2. In particular, u is a maximal element in $\mathbf{y}$. Let w be any element of $\mathbf{r}_q$ with $u \neq w$. If $w \in P_2$, then $w \in \mathbf{y}$, so $u \not< w$. If $w \in P_1$, then by properties of

the decomposition of P into $P_1 \triangleleft P_2$, it cannot be the case that $u < w$. Therefore, u is a maximal element of $\mathbf{r}_q$ of color γ. But this contradicts the fact that $\mathbf{r}_q$ is the minimal element in $\mathbf{comp}_\gamma(\mathbf{s})$. So it is not the case that $\mathbf{r}^{(2)} < \mathbf{y}$. Therefore, $\mathbf{r}^{(2)} = \mathbf{y}$, and so $\rho_\gamma^{(2)}(\mathbf{s} \cap P_2) = q'$. Combine this with $\rho_\gamma^{(1)}(\mathbf{s} \cap P_1) = q - q'$ to see that $\rho_\gamma(\mathbf{s}) = q = (q - q') + q' = \rho_\gamma^{(1)}(\mathbf{s} \cap P_1) + \rho_\gamma^{(2)}(\mathbf{s} \cap P_2)$.

The dual P^* may be viewed as a two-color grid poset that decomposes into $P_2^* \triangleleft P_1^*$. Order ideals of P^* are complements of order ideals of P. Then arguments analogous to those above apply to the complements of elements of the sequence $\mathbf{s} = \mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_p$. So we obtain $\rho_\gamma^*(P \setminus \mathbf{s}) = \rho_\gamma^{*(2)}((P \setminus \mathbf{s}) \cap P_2) + \rho_\gamma^{*(1)}((P \setminus \mathbf{s}) \cap P_1)$. Note that $(P \setminus \mathbf{s}) \cap P_i = P_i \setminus (\mathbf{s} \cap P_i)$ for $i \in \{1, 2\}$. Now $l_\gamma(\mathbf{s}) = \rho_\gamma(\mathbf{s}) + \rho_\gamma^*(P \setminus \mathbf{s})$, $l_\gamma^{(1)}(\mathbf{s} \cap P_1) = \rho_\gamma^{(1)}(\mathbf{s} \cap P_1) + \rho_\gamma^{*(1)}(P_1 \setminus (\mathbf{s} \cap P_1))$, and $l_\gamma^{(2)}(\mathbf{s} \cap P_2) = \rho_\gamma^{(2)}(\mathbf{s} \cap P_2) + \rho_\gamma^{*(2)}(P_2 \setminus (\mathbf{s} \cap P_2))$. Therefore, $l_\gamma(\mathbf{s}) = l_\gamma^{(1)}(\mathbf{s} \cap P_1) + l_\gamma^{(2)}(\mathbf{s} \cap P_2)$. □

It can be shown that if either of the conditions on the maximal and minimal elements on P_1 and P_2 required for the statement "$P = P_1 \triangleleft P_2$" fail, then so does at least one of the decomposition equations in Lemma 3.1 for $\rho_\gamma(\mathbf{s})$ and $l_\gamma(\mathbf{s})$.

4. $\mathfrak{g}$-semistandard posets, lattices, and tableaux. We define special two-color grid posets P as the "$\mathfrak{g}$-semistandard" posets. Then we define corresponding lattices $L = J_{color}(P)$ as the "$\mathfrak{g}$-semistandard" lattices. In the second half of the section, "$\mathfrak{g}$-semistandard" tableau descriptions of the elements of these lattices are developed.

For the remainder of this paper, $\mathfrak{g}$ denotes a rank two semisimple Lie algebra: $\mathfrak{g} \in \{A_1 \oplus A_1, A_2, C_2, G_2\}$. We identify α with a short simple root for $\mathfrak{g}$ and β as the other simple root. The vertex colors for the posets and the edge colors for the lattices which we now introduce correspond to the simple roots of $\mathfrak{g}$. So here the index set I of section 2 becomes $I = \{\alpha, \beta\}$. Let $\omega_\alpha = \omega_1 = (1, 0)$ and $\omega_\beta = \omega_2 = (0, 1)$, respectively, denote the corresponding fundamental weights. Then any weight μ in Λ of the form $\mu = p\omega_\alpha + q\omega_\beta$ (where p and q are integers) is now identified with the pair (p, q) in $\mathbb{Z} \times \mathbb{Z}$. In particular, α and β are, respectively, identified with the first and second row vectors from the Cartan matrix M for $\mathfrak{g}$. These matrices, displayed in Figure 4.1, specify the $\mathfrak{g}$-structure condition of section 2 for edge-colored ranked posets.

The $\mathfrak{g}$-*fundamental posets* $P_\mathfrak{g}(1, 0)$ and $P_\mathfrak{g}(0, 1)$ are defined to be the two-color grid posets of Figure 4.2. The corresponding $\mathfrak{g}$-*fundamental lattices* are defined to be the edge-colored lattices $L_\mathfrak{g}(1, 0) := J_{color}(P_\mathfrak{g}(1, 0))$ and $L_\mathfrak{g}(0, 1) := J_{color}(P_\mathfrak{g}(0, 1))$. See Figure 4.3. For the remainder of this section, everything presented for the *simple* cases (A_2, C_2, and G_2) has an easy $A_1 \oplus A_1$ analogue. The details for $A_1 \oplus A_1$ are omitted to save space, beginning with Figure 4.3.

Let $\lambda = (a, b)$, with $a, b \geq 0$. The $\mathfrak{g}$-*semistandard poset* $P_\mathfrak{g}^{\beta\alpha}(\lambda)$ associated to λ is defined to be the two-color grid poset P which has the decomposition $P_1 \triangleleft P_2 \triangleleft \cdots \triangleleft P_{a+b}$,

$A_1 \oplus A_1$	A_2	C_2	G_2
$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 \\ -2 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 \\ -3 & 2 \end{pmatrix}$

FIG. 4.1.

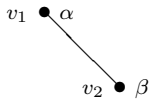
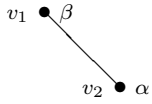
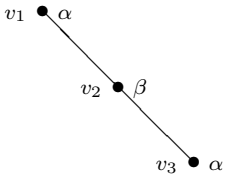
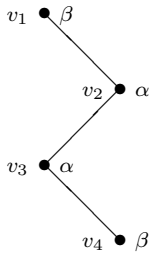
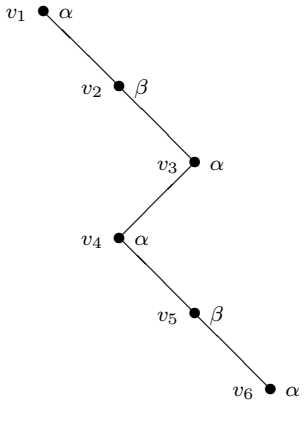
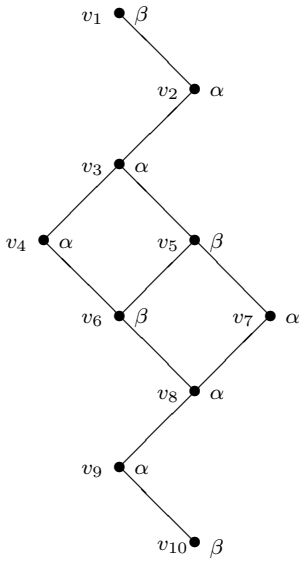
Algebra $\mathfrak{g}$	$P_{\mathfrak{g}}(1,0)$	$P_{\mathfrak{g}}(0,1)$
$A_1 \oplus A_1$		
A_2		
C_2		
G_2		

FIG. 4.2. $\mathfrak{g}$-*fundamental posets.*

where P_i is vertex-color isomorphic to $P_{\mathfrak{g}}(0,1)$ for $1 \leq i \leq b$ and to $P_{\mathfrak{g}}(1,0)$ for $1+b \leq i \leq a+b$. It can be seen that P is unique up to isomorphism. For each semisimple Lie algebra $\mathfrak{g}$, the poset $P_{\mathfrak{g}}^{\beta\alpha}(2,2)$ is depicted in Figure 3.2. The $\mathfrak{g}$-*semistandard poset* $P_{\mathfrak{g}}^{\alpha\beta}(\lambda)$ associated to λ is analogously defined, except with P_i vertex-color isomorphic to $P_{\mathfrak{g}}(1,0)$ for $1 \leq i \leq a$ and to $P_{\mathfrak{g}}(0,1)$ for $a+1 \leq i \leq a+b$. See Figure 3.3 for the corresponding $P_{\mathfrak{g}}^{\alpha\beta}(2,2)$. Note that $P_{\mathfrak{g}}^{\beta\alpha}(1,0) = P_{\mathfrak{g}}^{\alpha\beta}(1,0) = P_{\mathfrak{g}}(1,0)$, and $P_{\mathfrak{g}}^{\beta\alpha}(0,1) = P_{\mathfrak{g}}^{\alpha\beta}(0,1) = P_{\mathfrak{g}}(0,1)$. If $a = b = 0$, then $P_{\mathfrak{g}}^{\beta\alpha}(\lambda)$ and $P_{\mathfrak{g}}^{\alpha\beta}(\lambda)$ are the empty set. The $\mathfrak{g}$-*semistandard lattices* associated to λ are the edge-colored lattices

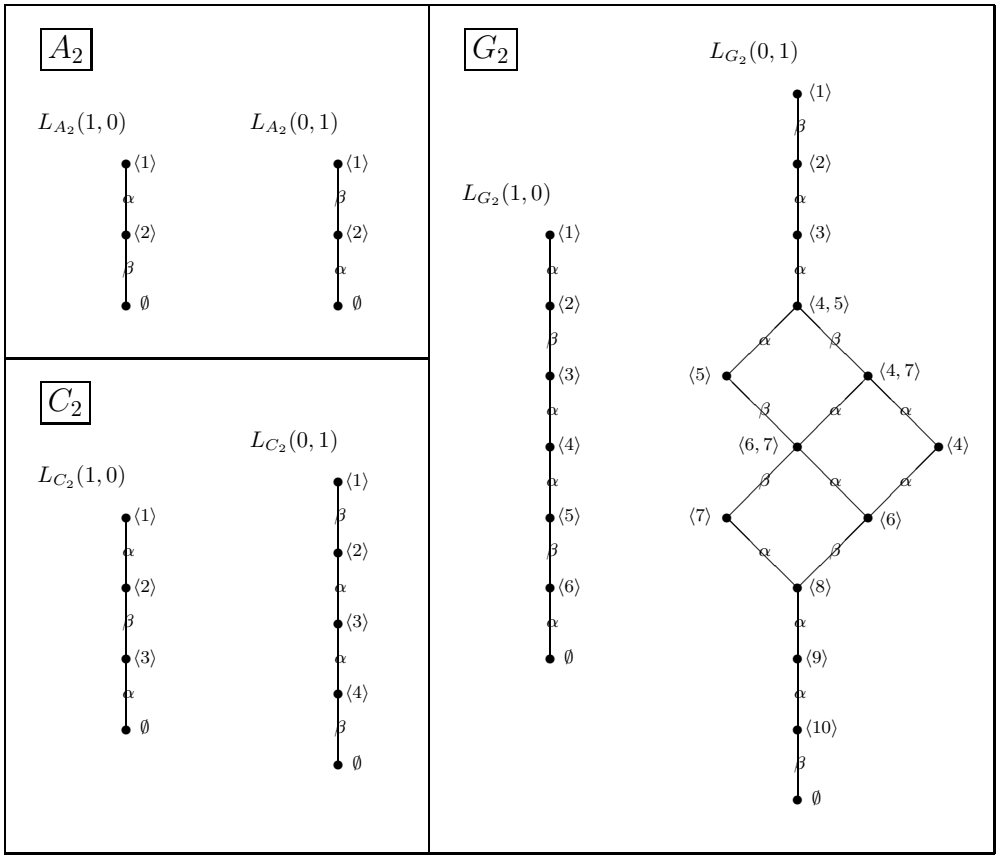


FIG. 4.3. *Elements of* $\mathfrak{g}$*-fundamental lattices as order ideals of* $\mathfrak{g}$*-fundamental posets. (Each order ideal is identified by the indices of its maximal vertices.)*

$L_{\mathfrak{g}}^{\beta\alpha}(\lambda) := J_{color}(P_{\mathfrak{g}}^{\beta\alpha}(\lambda))$ and $L_{\mathfrak{g}}^{\alpha\beta}(\lambda) := J_{color}(P_{\mathfrak{g}}^{\alpha\beta}(\lambda))$. Note that $L_{\mathfrak{g}}^{\beta\alpha}(1,0) = L_{\mathfrak{g}}^{\alpha\beta}(1,0) = L_{\mathfrak{g}}(1,0)$, and $L_{\mathfrak{g}}^{\beta\alpha}(0,1) = L_{\mathfrak{g}}^{\alpha\beta}(0,1) = L_{\mathfrak{g}}(0,1)$. We will not consider "mixed" concatenations, where some copies of $P_{\mathfrak{g}}(0,1)$ are interlaced amongst copies of $P_{\mathfrak{g}}(1,0)$. Any such concatenation will not have the max property, which *is* possessed by all of the $\mathfrak{g}$-semistandard posets.

Each $\mathfrak{g}$-semistandard lattice is an edge-colored poset. From now on we write $wt(\mathbf{s})$ for $wt_L(\mathbf{s})$ when L is $\mathfrak{g}$-semistandard. Let $\mathbf{s} \in L$. Let $\gamma \in \{\alpha, \beta\}$. By definition, the γ-entry of the 2-tuple $wt(\mathbf{s})$ is the rank of $\mathbf{s}$ within the γ-colored connected component of $\mathbf{s}$ diminished by the depth of $\mathbf{s}$ in that component.

LEMMA 4.1. *Let* $\mathbf{s} \xrightarrow{\gamma} \mathbf{t}$ *be an edge of color* $\gamma \in \{\alpha, \beta\}$ *in a* $\mathfrak{g}$*-fundamental lattice* L*. Then* $wt(\mathbf{s}) + \gamma = wt(\mathbf{t})$*. Hence each* $\mathfrak{g}$*-fundamental lattice satisfies the* $\mathfrak{g}$*-structure condition.*

Proof. Note that $\mathbf{s}$ and $\mathbf{t}$ are in the same γ-component. Since $\mathbf{t}$ covers $\mathbf{s}$ in this component, the γ-entry of $wt(\mathbf{t})$ is 2 more than the γ-entry of $wt(\mathbf{s})$. But adding the simple root γ to $wt(\mathbf{s})$ adds 2 to the γ-entry of $wt(\mathbf{s})$, since $M_{\gamma,\gamma} = 2$ always. Let γ' in I be such that $\gamma' \neq \gamma$. Using Figure 4.4, one can quickly check by hand that the γ'-entry of $wt(\mathbf{s})$ changes by $M_{\gamma,\gamma'}$ or by $M_{\gamma',\gamma}$ (as appropriate) for each edge within each γ-component of a $\mathfrak{g}$-fundamental lattice. ∎

PROPOSITION 4.2. *Let $\lambda = (a, b)$, with $a, b \geq 0$. Let L be one of the $\mathfrak{g}$-semistandard lattices $L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$ or $L_{\mathfrak{g}}^{\alpha\beta}(\lambda)$. Let $\mathbf{s} \xrightarrow{\gamma} \mathbf{t}$ be an edge of color $\gamma \in \{\alpha, \beta\}$ in L. Then $wt(\mathbf{s}) + \gamma = wt(\mathbf{t})$, and hence L satisfies the $\mathfrak{g}$-structure condition.*

Proof. In light of Lemma 4.1, apply part (2) of Lemma 3.1. □

Remark 4.3. If $\mathfrak{g} = A_1 \oplus A_1$, then we have $P_{\mathfrak{g}}^{\beta\alpha}(\lambda) \cong P_{\mathfrak{g}}^{\alpha\beta}(\lambda)$ as vertex-colored posets: their Hasse diagrams are vertex-color isomorphic to $P_{\mathfrak{g}}^{\beta\alpha}(a, 0) \oplus P_{\mathfrak{g}}^{\beta\alpha}(0, b) \cong \mathbf{a} \oplus \mathbf{b}$. Hence $L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$ and $L_{\mathfrak{g}}^{\alpha\beta}(\lambda)$ are edge-color isomorphic to $L_{\mathfrak{g}}^{\beta\alpha}(a, 0) \times L_{\mathfrak{g}}^{\beta\alpha}(0, b) \cong (\mathbf{a} + \mathbf{1}) \times (\mathbf{b} + \mathbf{1})$. For $\mathfrak{g} = C_2$ or $\mathfrak{g} = G_2$, observe that $P_{\mathfrak{g}}^{\alpha\beta}(\lambda)$ is vertex-color isomorphic to $(P_{\mathfrak{g}}^{\beta\alpha}(\lambda))^*$, and thus $L_{\mathfrak{g}}^{\alpha\beta}(\lambda)$ and $(L_{\mathfrak{g}}^{\beta\alpha}(\lambda))^*$ are isomorphic as edge-colored posets. For $\mathfrak{g} = A_2$, $P_{\mathfrak{g}}^{\alpha\beta}(\lambda)$ and $(P_{\mathfrak{g}}^{\beta\alpha}(\lambda))^*$ are isomorphic as posets, but their vertex colors are reversed; disregarding edge colors, it follows that $L_{\mathfrak{g}}^{\alpha\beta}(\lambda)$ and $(L_{\mathfrak{g}}^{\beta\alpha}(\lambda))^*$ are isomorphic as posets. In all cases, $L_{\mathfrak{g}}^{\alpha\beta}(\lambda) \cong (L_{\mathfrak{g}}^{\beta\alpha}(\lambda))^{\triangle}$.

The easy proof of the following statement will be omitted.

LEMMA 4.4. *Let $\lambda = (a, b)$, with $a, b \geq 0$. If $\mathfrak{g}$ is simple, then $L_{\mathfrak{g}}^{\beta\alpha}(\lambda) \cong L_{\mathfrak{g}}^{\alpha\beta}(\lambda)$ as edge-colored posets if and only if $a = 0$ or $b = 0$.*

Now we develop tableau labels for the elements of half of the $\mathfrak{g}$-semistandard lattices, the $L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$. Comments relating these tableaux to tableaux developed by some of us and other authors appear in section 5. We associate to the fundamental weight $\omega_\alpha = (1, 0)$ the shape $\mathbf{shape}(1, 0) = \ydiagram{1}$; we associate to $\omega_\beta = (0, 1)$ the shape $\mathbf{shape}(0, 1) = \ydiagram{1,1}$. For $a, b \geq 0$, we associate to $\lambda = (a, b)$ the shape (Ferrers diagram) with b columns of length two and a columns of length one. A *tableau of shape* λ is a filling of the boxes of $\mathbf{shape}(\lambda)$ with entries from some totally ordered set. For a tableau T of shape λ, we write $T = (T^{(1)}, \ldots, T^{(a+b)})$, where $T^{(i)}$ is the ith column of T from the left. We let $T_j^{(i)}$ denote the jth entry of the column $T^{(i)}$, counting from the top. The tableau T is *semistandard* if the entries weakly increase across rows and strictly increase down columns. To each element $\mathbf{t}$ of a $\mathfrak{g}$-fundamental lattice from Figure 4.3 we associate the one-column semistandard tableau $\mathbf{tableau}(\mathbf{t})$ of Figure 4.4. For an order ideal $\mathbf{t}$ of $P_{\mathfrak{g}}^{\beta\alpha}(\lambda)$, let $\mathbf{tableau}(\mathbf{t})$ be the tableau $T = (T^{(1)}, \ldots, T^{(a+b)})$ with $T^{(i)} = \mathbf{tableau}(\mathbf{t} \cap P_i)$. A tableau T of shape λ obtained in this way is a $\mathfrak{g}$-*semistandard tableau of shape* λ. We let $\mathcal{S}_{\mathfrak{g}}(\lambda)$ denote the set of all $\mathfrak{g}$-semistandard tableaux of shape λ. The function $\mathbf{tableau} : L_{\mathfrak{g}}^{\beta\alpha}(\lambda) \longrightarrow \mathcal{S}_{\mathfrak{g}}(\lambda)$ is a one-to-one correspondence. See Figure 6.1 for a C_2 example.

PROPOSITION 4.5. *Let $a, b \geq 0$, and let $\lambda = (a, b)$. Then*

$$\mathcal{S}_{A_2}(\lambda) = \Big\{ \textit{semistandard tableau } T \textit{ of shape} \lambda \textit{ with entries from } \{1, 2, 3\} \Big\},$$

$$\mathcal{S}_{C_2}(\lambda) = \Big\{ \textit{semistandard tableau } T \textit{ of shape } \lambda \textit{ with entries from } \{1, 2, 3, 4\} \;\Big|\; \ytableaushort{14} \textit{ is not a column of } T, \textit{ and } \ytableaushort{23} \textit{ appears at most once in } T \Big\}$$

$$\mathcal{S}_{G_2}(\lambda) = \Big\{ \textit{semistandard tableau } T \textit{ of shape } \lambda \textit{ with entries from } \{1, 2, 3, 4, 5, 6, 7\} \;\Big|\; \textit{the column } \ytableaushort{4} \textit{ appears at most once in } T; \ytableaushort{23}, \ytableaushort{24}, \ytableaushort{34}, \ytableaushort{35}, \ytableaushort{45}, \ytableaushort{46}, \textit{and } \ytableaushort{56} \textit{ are not columns of } T; \textit{ plus the restrictions of Figure 4.5} \Big\}.$$

Proof. The association of one-column $\mathfrak{g}$-semistandard tableaux with order ideals of $\mathfrak{g}$-fundamental posets is given in Figures 4.3 and 4.4. Consider the $\mathfrak{g} = C_2$ case.

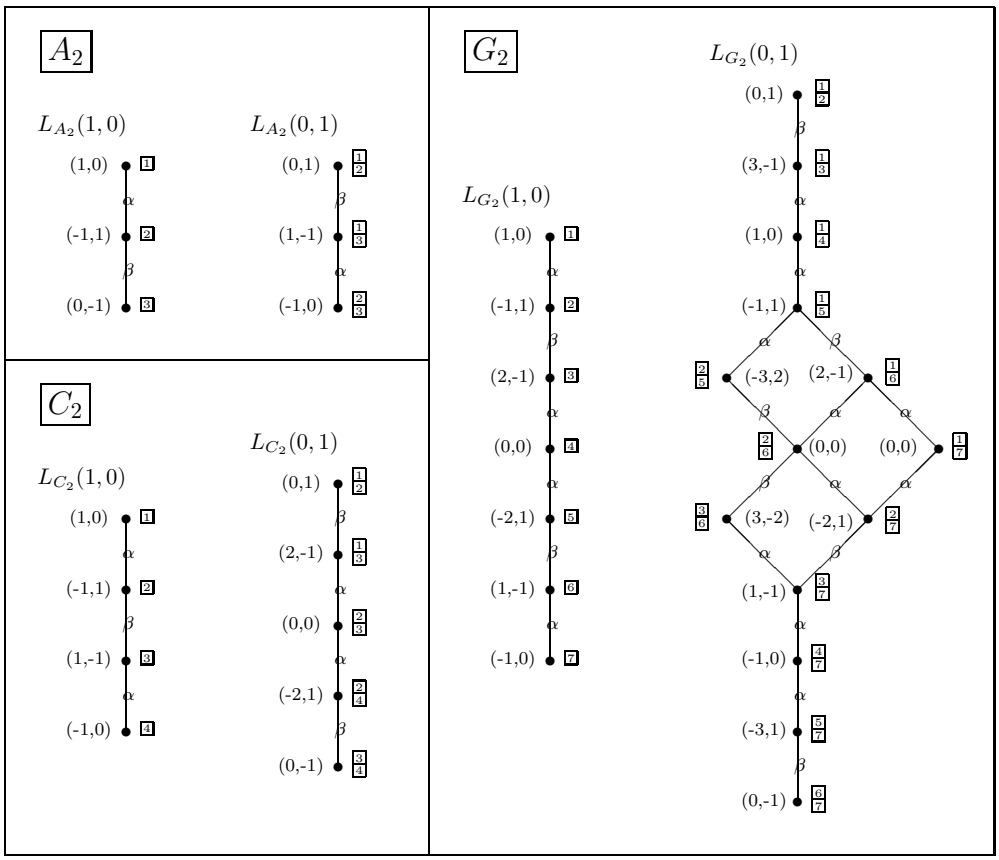


FIG. 4.4. *Weights and tableaux for* $\mathfrak{g}$*-fundamental lattices.*

We want to show that the set $\mathcal{S}_{C_2}(\lambda)$ is the same as the stated set, which we denote $\mathcal{S}$. Let $T \in \mathcal{S}_{C_2}(\lambda)$, so $T = \mathbf{tableau}(\mathbf{t})$ for some order ideal $\mathbf{t}$ of $P_{C_2}^{\beta\alpha}(\lambda)$. Write $P_{C_2}^{\beta\alpha}(\lambda) = P_1 \lhd \cdots \lhd P_{a+b}$, as depicted in Figure 4.6. Following Figure 4.2, we label the vertices of P_j as $w_{1,j}$, $w_{2,j}$, $w_{3,j}$, and $w_{4,j}$ with $w_{1,j} > w_{2,j} > w_{3,j} > w_{4,j}$ whenever $1 \le j \le b$, and we label the vertices of P_j as $z_{1,j}$, $z_{2,j}$, and $z_{3,j}$ with $z_{1,j} > z_{2,j} > z_{3,j}$ whenever $1 + b \le j \le a + b$. By definition, $T = (T^{(1)}, \ldots, T^{(a+b)})$ with $T^{(i)} = \mathbf{tableau}(\mathbf{t} \cap P_i)$. The entries for $T^{(i)}$ are from the set $\{1, 2, 3, 4\}$, and no $T^{(i)}$ is the column $\begin{array}{|c|}\hline 1 \\ \hline 4 \\ \hline \end{array}$. To see how the semistandard and other restrictions occur, suppose (for example) that $T^{(i)}$ is the column $\begin{array}{|c|}\hline 2 \\ \hline 3 \\ \hline \end{array}$ for some $1 \le i \le b$. Note that $\mathbf{t} \cap P_i = \{w_{3,i}, w_{4,i}\}$. It follows that $w_{1,j}$, $w_{2,j}$, and $w_{3,j}$ are not in $\mathbf{t}$ for $i < j \le b$, and moreover $z_{1,j}$ is not in $\mathbf{t}$ for $1 + b \le j \le a + b$. In particular, it follows that $T^{(i+1)}$ cannot be $\begin{array}{|c|}\hline 1 \\ \hline 2 \\ \hline \end{array}$, $\begin{array}{|c|}\hline 1 \\ \hline 3 \\ \hline \end{array}$, $\begin{array}{|c|}\hline 2 \\ \hline 3 \\ \hline \end{array}$, or $\begin{array}{|c|}\hline 1 \\ \hline \end{array}$, assuming the column $T^{(i+1)}$ exists. That is, the pair of columns $T^{(i)}$ and $T^{(i+1)}$ meets the requirements for inclusion in the set $\mathcal{S}$. The other eight cases for $T^{(i)}$ can be handled in a similar fashion. We conclude that $T \in \mathcal{S}$. So $\mathcal{S}_{C_2}(\lambda) \subseteq \mathcal{S}$.

In the other direction, suppose $T = (T^{(1)}, \ldots, T^{(a+b)})$ is in $\mathcal{S}$. For each i, let Q_i be the order ideal of P_i corresponding to the one-column tableau $T^{(i)}$, and let $\mathbf{t} := \cup_i Q_i$. By examining cases as in the previous paragraph, one can check that the

Column $T^{(i)}$ of T	Then the succeeding column $T^{(i+1)}$ of T cannot be...
$\boxed{4}$	$\boxed{4}$
$\boxed{\begin{smallmatrix}1\\4\end{smallmatrix}}$	$\boxed{1}$, $\boxed{\begin{smallmatrix}1\\4\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}1\\5\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}1\\6\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}1\\7\end{smallmatrix}}$
$\boxed{\begin{smallmatrix}1\\5\end{smallmatrix}}$	$\boxed{1}$, $\boxed{\begin{smallmatrix}1\\5\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}1\\6\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}1\\7\end{smallmatrix}}$
$\boxed{\begin{smallmatrix}1\\6\end{smallmatrix}}$	$\boxed{1}$, $\boxed{2}$, $\boxed{\begin{smallmatrix}1\\6\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}1\\7\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}2\\6\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}2\\7\end{smallmatrix}}$
$\boxed{\begin{smallmatrix}2\\6\end{smallmatrix}}$	$\boxed{2}$, $\boxed{\begin{smallmatrix}2\\6\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}2\\7\end{smallmatrix}}$
$\boxed{\begin{smallmatrix}1\\7\end{smallmatrix}}$	$\boxed{1}$, $\boxed{2}$, $\boxed{3}$, $\boxed{4}$, $\boxed{\begin{smallmatrix}1\\7\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}2\\7\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}3\\7\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}4\\7\end{smallmatrix}}$
$\boxed{\begin{smallmatrix}2\\7\end{smallmatrix}}$	$\boxed{2}$, $\boxed{3}$, $\boxed{4}$, $\boxed{\begin{smallmatrix}2\\7\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}3\\7\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}4\\7\end{smallmatrix}}$
$\boxed{\begin{smallmatrix}3\\7\end{smallmatrix}}$	$\boxed{3}$, $\boxed{4}$, $\boxed{\begin{smallmatrix}3\\7\end{smallmatrix}}$, $\boxed{\begin{smallmatrix}4\\7\end{smallmatrix}}$
$\boxed{\begin{smallmatrix}4\\7\end{smallmatrix}}$	$\boxed{4}$, $\boxed{\begin{smallmatrix}4\\7\end{smallmatrix}}$

FIG. 4.5. *Some restrictions for any given G_2-semistandard tableau T.*

restrictions on T as an element of $\mathcal{S}$ guarantee that $\mathbf{t}$ will be an order ideal of $P_{C_2}^{\beta\alpha}(\lambda)$ with $Q_i = \mathbf{t} \cap P_i$ for each i. Hence $T \in \mathcal{S}_{C_2}(\lambda)$. It follows that $\mathcal{S} \subseteq \mathcal{S}_{C_2}(\lambda)$, which completes the proof for the C_2 case. The A_2 and G_2 cases can be handled by similar arguments. ∎

Remark 4.6. In passing we note that the partial ordering and the covering relations in $L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$ are easy to describe with the "coordinates" of $\mathfrak{g}$-semistandard tableaux. For $\mathbf{s}$ and $\mathbf{t}$ in $L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$, let $S := \mathbf{tableau}(\mathbf{s})$ and $T := \mathbf{tableau}(\mathbf{t})$. Then $\mathbf{s} \leq \mathbf{t}$ if and only if $S_j^{(i)} \geq T_j^{(i)}$ for all i, j. (This is the "reverse componentwise" order on tableaux.) Moreover, $\mathbf{s} \to \mathbf{t}$ is a covering relation in the poset $L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$ if and only if for some i and j we have $S_j^{(i)} = T_j^{(i)} + 1$ while $S_q^{(p)} = T_q^{(p)}$ for all $(p, q) \neq (i, j)$. For $\mathfrak{g} = A_2$, the edge gets color α if $T_j^{(i)}$ is 1 and color β if $T_j^{(i)}$ is 2; for $\mathfrak{g} = C_2$, the edge gets color α if $T_j^{(i)}$ is 1 or 3 and color β if $T_j^{(i)}$ is 2; for $\mathfrak{g} = G_2$, the edge gets color α if $T_j^{(i)}$ is 1 or 3 or 4 or 6 and color β if $T_j^{(i)}$ is 2 or 5.

For a tableau T of shape $\lambda = (a, b)$ and a positive integer k, we define $n_k(T)$ to be the number of times the entry k appears in the tableau T. Observe that $n_k(T) = \sum_{i=1}^{a+b} n_k(T^{(i)})$. Define a function $\mathbf{tableauwt} : \mathcal{S}_{\mathfrak{g}}(\lambda) \longrightarrow \mathbb{Z} \times \mathbb{Z}$ by the rules:

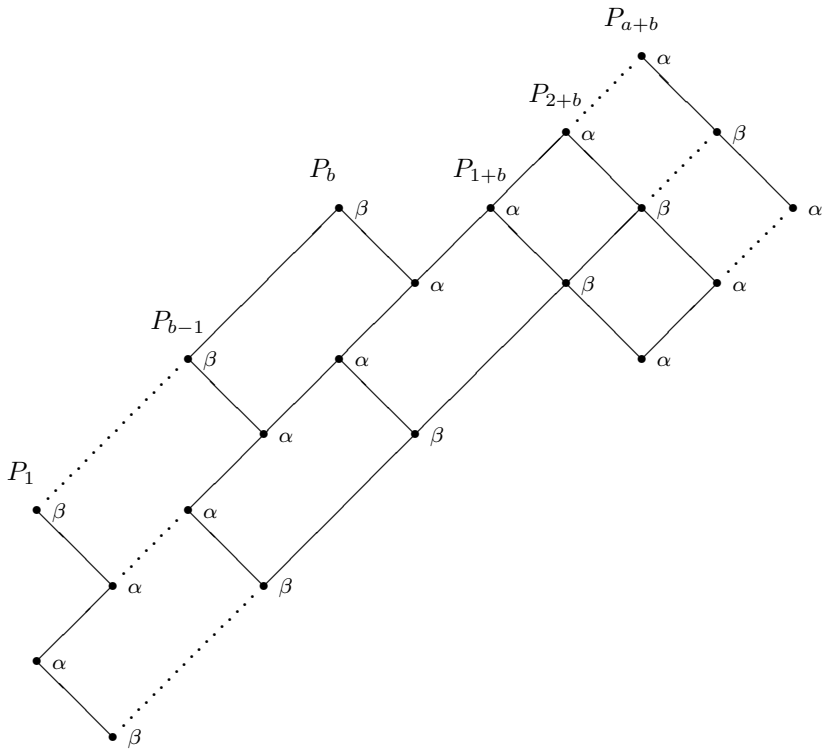


FIG. 4.6. $P_{C_2}^{\beta\alpha}(\lambda) = P_1 \lhd \cdots \lhd P_{a+b}$.

$$\mathbf{tableauwt}(T) := \begin{cases} \Big(n_1(T) - n_2(T), n_2(T) - n_3(T)\Big) & \text{if } \mathfrak{g} = A_2, \\ \Big(n_1(T) - n_2(T) + n_3(T) - n_4(T), n_2(T) - n_3(T)\Big) & \text{if } \mathfrak{g} = C_2, \\ \Big(n_1(T) - n_2(T) + 2n_3(T) - 2n_5(T) + n_6(T) - n_7(T), & \\ \qquad n_2(T) - n_3(T) + n_5(T) - n_6(T)\Big) & \text{if } \mathfrak{g} = G_2. \end{cases}$$

The function $wt(\mathbf{s})$ defined on $L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$ in terms of the color components of $L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$ can be expressed in terms of the tableau entry counts when the elements $\mathbf{s}$ of $L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$ are viewed as tableaux $\mathbf{t}$ in $\mathcal{S}_{\mathfrak{g}}(\lambda)$.

PROPOSITION 4.7. *Let $\lambda = (a, b)$, with $a, b \geq 0$. For $\mathbf{t} \in L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$, consider $T := \mathbf{tableau}(\mathbf{t}) \in \mathcal{S}_{\mathfrak{g}}(\lambda)$. Then $wt(\mathbf{t}) = \mathbf{tableauwt}(T)$.*

Proof. With the help of Figures 4.4 and 4.5, one can easily confirm the result by hand whenever λ is a fundamental weight. Then, more generally, one can apply Lemma 3.1 to $wt(\mathbf{t})$, noting that $\mathbf{tableauwt}(T) = \sum_{i=1}^{a+b} \mathbf{tableauwt}(T^{(i)})$. ∎

This concludes our self-contained development of $\mathfrak{g}$-semistandard posets, $\mathfrak{g}$-semistandard lattices, and $\mathfrak{g}$-semistandard tableaux in sections 3 and 4.

5. Weyl characters; Littelmann's tableaux; main results. Our main result, Theorem 5.3, expresses the Weyl characters for the irreducible representations of the rank two semisimple Lie algebras as generating functions for $\mathfrak{g}$-semistandard lat-

Algebra	Simple roots	Positive roots	Weyl group W (By generators and relations; as reduced words)
A_2	$\alpha = 2\omega_\alpha - \omega_\beta$ $\beta = -\omega_\alpha + 2\omega_\beta$	$\alpha, \beta, \alpha+\beta$	$\langle s_\alpha, s_\beta \mid s_\alpha^2 = s_\beta^2 = id, (s_\alpha s_\beta)^3 = id\rangle$ $\{id, s_\alpha, s_\beta, s_\alpha s_\beta, s_\beta s_\alpha, s_\alpha s_\beta s_\alpha = s_\beta s_\alpha s_\beta\}$
C_2	$\alpha = 2\omega_\alpha - \omega_\beta$ $\beta = -2\omega_\alpha + 2\omega_\beta$	$\alpha, \beta, \alpha+\beta,$ $2\alpha+\beta$	$\langle s_\alpha, s_\beta \mid s_\alpha^2 = s_\beta^2 = id, (s_\alpha s_\beta)^4 = id\rangle$ $\{id, s_\alpha, s_\beta, s_\alpha s_\beta, s_\beta s_\alpha, s_\alpha s_\beta s_\alpha, s_\beta s_\alpha s_\beta,$ $s_\alpha s_\beta s_\alpha s_\beta = s_\beta s_\alpha s_\beta s_\alpha\}$
G_2	$\alpha = 2\omega_\alpha - \omega_\beta$ $\beta = -3\omega_\alpha + 2\omega_\beta$	$\alpha, \beta, \alpha+\beta,$ $2\alpha+\beta,$ $3\alpha+\beta,$ $3\alpha+2\beta$	$\langle s_\alpha, s_\beta \mid s_\alpha^2 = s_\beta^2 = id, (s_\alpha s_\beta)^6 = id\rangle$ $\{id, s_\alpha, s_\beta, s_\alpha s_\beta, s_\beta s_\alpha, s_\alpha s_\beta s_\alpha, s_\beta s_\alpha s_\beta,$ $s_\alpha s_\beta s_\alpha s_\beta, s_\beta s_\alpha s_\beta s_\alpha, s_\alpha s_\beta s_\alpha s_\beta s_\alpha,$ $s_\beta s_\alpha s_\beta s_\alpha s_\beta, s_\alpha s_\beta s_\alpha s_\beta s_\alpha s_\beta = s_\beta s_\alpha s_\beta s_\alpha s_\beta s_\alpha\}$

FIG. 5.1. *Roots and Weyl groups for the rank two simple Lie algebras.*

tices. We begin by recording some explicit data on roots, weights, Weyl groups, and irreducible characters for the rank two semisimple Lie algebras. Then we describe certain tableaux obtained by Littelmann in [Lit], and in Proposition 5.2 we match these with our $\mathfrak{g}$-semistandard tableaux. Corollary 5.4 gives the product expressions for the rank generating functions.

In rank two we denote the elements e_{ω_α} and e_{ω_β} of the group ring $\mathbb{Z}[\Lambda]$ by x and y. Then, in the notation of section 2, the irreducible Weyl character χ_λ for $\mathfrak{g}$ is a Laurent polynomial in the variables x and y denoted $char_{\mathfrak{g}}(\lambda; x, y)$. The reflections s_α and s_β in W act on the fundamental weights as follows: $s_\alpha\omega_\alpha = \omega_\alpha - \alpha$, $s_\alpha\omega_\beta = \omega_\beta$, $s_\beta\omega_\alpha = \omega_\alpha$, and $s_\beta\omega_\beta = \omega_\beta - \beta$. Figure 5.1 has data for the simple roots, positive roots, and Weyl group for each of the rank two simple Lie algebras. Recall from section 2 that the denominator A_ϱ of the Weyl character formula can be expressed as a product over the positive roots. Also recall that the numerator $A_{\varrho+\lambda}$ is an alternating sum over the elements of the Weyl group. Using the data of Figure 5.1 one obtains for A_2

$$\begin{aligned} A_\varrho &= xy(1-x^{-2}y)(1-xy^{-2})(1-x^{-1}y^{-1}) \\ &= xy - x^{-1}y^2 - x^2y^{-1} + x^{-2}y + xy^{-2} - x^{-1}y^{-1}, \\ A_{\varrho+\lambda} &= x^{a+1}y^{b+1} - x^{-(a+1)}y^{a+b+2} - x^{a+b+2}y^{-(b+1)} \\ &\quad + x^{-(a+b+2)}y^{a+1} + x^{b+1}y^{-(a+b+2)} - x^{-(b+1)}y^{-(a+1)}. \end{aligned}$$

For C_2 we get

$$\begin{aligned} A_\varrho &= xy(1-x^{-2}y)(1-x^2y^{-2})(1-y^{-1})(1-x^{-2}) \\ &= xy - x^{-1}y^2 - x^3y^{-1} + x^{-3}y^2 + x^3y^{-2} - x^{-3}y^1 - xy^{-2} + x^{-1}y^{-1}, \\ A_{\varrho+\lambda} &= x^{a+1}y^{b+1} - x^{-(a+1)}y^{a+b+2} - x^{a+2b+3}y^{-(b+1)} + x^{-(a+2b+3)}y^{a+b+2} \\ &\quad + x^{a+2b+3}y^{-(a+b+2)} - x^{-(a+2b+3)}y^{b+1} - x^{a+1}y^{-(a+b+2)} + x^{-(a+1)}y^{-(b+1)}. \end{aligned}$$

And for G_2 we have

$$\begin{aligned}
A_\varrho &= xy(1-x^{-2}y)(1-x^3y^{-2})(1-xy^{-1})(1-x^{-1})(1-x^{-3}y)(1-y^{-1})\\
&= xy - x^{-1}y^2 - x^4y^{-1} + x^{-4}y^3 + x^5y^{-2} - x^{-5}y^3 - x^5y^{-3} + x^{-5}y^2 + x^4y^{-3}\\
&\quad - x^{-4}y - xy^{-2} + x^{-1}y^{-1},\\
A_{\varrho+\lambda} &= x^{a+1}y^{b+1} - x^{-(a+1)}y^{a+b+2} - x^{a+3b+4}y^{-(b+1)} + x^{-(a+3b+4)}y^{a+2b+3}\\
&\quad + x^{2a+3b+5}y^{-(a+b+2)}\\
&\quad - x^{-(2a+3b+5)}y^{a+2b+3} - x^{2a+3b+5}y^{-(a+2b+3)} + x^{-(2a+3b+5)}y^{a+b+2}\\
&\quad + x^{a+3b+4}y^{-(a+2b+3)}\\
&\quad - x^{-(a+3b+4)}y^{b+1} - x^{a+1}y^{-(a+b+2)} + x^{-(a+1)}y^{-(b+1)}.
\end{aligned}$$

We now seek a correspondence between our $\mathfrak{g}$-semistandard tableaux and certain tableaux of Littelmann [Lit]. Littelmann's tableaux are "translations" of the standard monomial theory tableaux of Lakshmibai and Seshadri. The roles of his columns and rows are reversed with respect to this paper. We preprocess Littelmann's tableaux in two steps. First, we reflect them across the main diagonal $i = j$. Then we group k of his columns at a time into a "block" of k columns, where $k = 1$ for A_2, $k = 2$ for C_2, and $k = 6$ for G_2. We define $\mathbf{shape}(k \times \lambda) := \mathbf{shape}(\mu)$, where $\mu = ka\omega_\alpha + kb\omega_\beta = (ka, kb)$. A *$k$-tableau of shape λ* is a filling of $\mathbf{shape}(k \times \lambda)$ with entries from some totally ordered set. The *semistandard* condition on k-tableaux is the same as the semistandard condition of section 4. For a k-tableau T of shape λ, we write $T = (T^{(1)}, \dots, T^{(a+b)})$, where $T^{(i)}$ is the ith block of k columns of T counting from the left. Only certain fillings of these k-column blocks will be "admissible." Here are our processed versions of Littelmann's tableaux.

DEFINITION 5.1. *Let $\lambda = (a, b)$, with $a, b \geq 0$. Then*

$$\begin{aligned}
\mathcal{LT}_{A_2}(\lambda) &:= \Big\{ \textit{semistandard 1-tableau } T \textit{ of shape } \lambda \textit{ with entries from } \{1,2,3\} \;\Big|\\
&\qquad \textit{admissible 1-column blocks of } T \textit{ come from Figure 5.2} \Big\},\\
\mathcal{LT}_{C_2}(\lambda) &:= \Big\{ \textit{semistandard 2-tableau } T \textit{ of shape } \lambda \textit{ with entries from } \{1,2,3,4\} \;\Big|\\
&\qquad \textit{admissible 2-column blocks of } T \textit{ come from Figure 5.2} \Big\},\\
\mathcal{LT}_{G_2}(\lambda) &:= \Big\{ \textit{semistandard 6-tableau } T \textit{ of shape } \lambda \textit{ with entries from } \{1,2,3,4,5,6\} \;\Big|\\
&\qquad \textit{admissible 6-column blocks of } T \textit{ come from Figure 5.2} \Big\}.
\end{aligned}$$

Moreover, the weight $wt_{Lit}(T)$ of a Littelmann tableau T is given by

$$wt_{Lit}(T) := \begin{cases}
\Big(n_1(T) - n_2(T)\Big)\omega_\alpha + \Big(n_2(T) - n_3(T)\Big)\omega_\beta & \text{if } \mathfrak{g} = A_2,\\
\frac{1}{2}\Big[\Big(n_1(T) - n_2(T) + n_3(T) - n_4(T)\Big)\omega_\alpha + \Big(n_2(T) - n_3(T)\Big)\omega_\beta\Big] & \text{if } \mathfrak{g} = C_2,\\
\frac{1}{6}\Big[\Big(n_1(T) - n_2(T) + 2n_3(T) - 2n_4(T) + n_5(T) - n_6(T)\Big)\omega_\alpha & \\
\qquad + \Big(n_2(T) - n_3(T) + n_4(T) - n_5(T)\Big)\omega_\beta\Big] & \text{if } \mathfrak{g} = G_2.
\end{cases}$$

To obtain these tableaux for A_2, see section 2 of [Lit]; for C_2, see the appendix of [Lit]; and for G_2 see section 3 of that paper. Littelmann expresses his weight function

in terms of a basis $\{\varepsilon_1, \varepsilon_2\}$ for Λ, where $\varepsilon_1 = \omega_\alpha$ and $\varepsilon_2 = \omega_\beta - \omega_\alpha$. A consequence of standard monomial theory is the following theorem.

THEOREM (Littelmann, Lakshmibai, Seshadri). *Let $\lambda = (a, b)$, with $a, b \geq 0$. Let $\mathfrak{g}$ be a rank two simple Lie algebra. Then $(\mathcal{LT}_{\mathfrak{g}}(\lambda), wt_{Lit})$ is a splitting system for the irreducible character χ_λ.*

Next we describe a weight-preserving bijection $\phi : \mathcal{S}_{\mathfrak{g}}(\lambda) \longrightarrow \mathcal{LT}_{\mathfrak{g}}(\lambda)$. For fundamental weights, the correspondence between the $\mathfrak{g}$-semistandard tableaux of section 4 and Littelmann k-tableaux of this section is given in Figure 5.2. Given a rank two simple Lie algebra $\mathfrak{g}$, a dominant weight $\lambda = a\omega_\alpha + b\omega_\beta = (a, b)$, and a tableau T in $\mathcal{S}_{\mathfrak{g}}(\lambda)$, we let $U = \phi(T)$ be the Littelmann k-tableau of shape λ whose ith k-column block $U^{(i)}$ corresponds to the ith column $T^{(i)}$ of T. Keeping in mind the restrictions on which columns can follow $T^{(i)}$ to form a $\mathfrak{g}$-semistandard tableau T in $\mathcal{S}_{\mathfrak{g}}(\lambda)$, one can check that $U^{(i)}$ followed by $U^{(i+1)}$ obeys the semistandard requirement for Littelmann k-tableaux. Hence U is in $\mathcal{LT}_{\mathfrak{g}}(\lambda)$. Similarly, given U in $\mathcal{LT}_{\mathfrak{g}}(\lambda)$, let $T = \psi(U)$ be the tableau of shape λ whose ith column $T^{(i)}$ corresponds to the ith k-column block $U^{(i)}$ of U. Keeping in mind the semistandard condition on the Littelmann k-tableaux in $\mathcal{LT}_{\mathfrak{g}}(\lambda)$, one can check that $T^{(i)}$ followed by $T^{(i+1)}$ obeys the restrictions for $\mathfrak{g}$-semistandard tableaux in $\mathcal{S}_{\mathfrak{g}}(\lambda)$, and hence T is in $\mathcal{S}_{\mathfrak{g}}(\lambda)$. Clearly the mappings ϕ and ψ are inverses.

PROPOSITION 5.2. *Keep the notation of the previous paragraph. The mapping $\phi : \mathcal{S}_{\mathfrak{g}}(\lambda) \longrightarrow \mathcal{LT}_{\mathfrak{g}}(\lambda)$ described above is a weight-preserving bijection: for any $T \in \mathcal{S}_{\mathfrak{g}}(\lambda)$, $wt_{Lit}(\phi(T)) = \mathbf{tableauwt}(T)$.*

Proof. We must check that ϕ is weight-preserving. If λ is a fundamental weight, simply inspect Figure 5.2. If λ is a dominant weight and T is in $\mathcal{S}_{\mathfrak{g}}(\lambda)$, then $\mathbf{tableauwt}(T) = \sum \mathbf{tableauwt}(T^{(i)}) = \sum wt_{Lit}(\phi(T^{(i)}))$. The characterization of wt_{Lit} in Definition 5.1 implies that $wt_{Lit}(\phi(T)) = \sum wt_{Lit}(\phi(T^{(i)}))$. ∎

THEOREM 5.3. *Let $\mathfrak{g}$ be a semisimple Lie algebra of rank two. Let $\lambda = (a, b)$, with $a, b \geq 0$. Let L be one of the $\mathfrak{g}$-semistandard lattices $L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$ or $L_{\mathfrak{g}}^{\alpha\beta}(\lambda)$. Then L is a splitting poset for an irreducible representation of $\mathfrak{g}$ with highest weight λ. In particular,*

$$char_{\mathfrak{g}}(\lambda; x, y) = \sum_{\mathbf{s} \in L} (x, y)^{wt(\mathbf{s})}.$$

Proof. Proposition 4.2 states that L satisfies the $\mathfrak{g}$-structure condition. Suppose $\mathfrak{g}$ is simple. Since $(\mathcal{LT}_{\mathfrak{g}}(\lambda), wt_{Lit})$ is a splitting system for χ_λ, it follows from Proposition 5.2 that $(\mathcal{S}_{\mathfrak{g}}(\lambda), \mathbf{tableauwt})$ is as well. From Proposition 4.7 it now follows that $(L_{\mathfrak{g}}^{\beta\alpha}(\lambda), wt)$ is a splitting system for χ_λ. Since $L_{\mathfrak{g}}^{\alpha\beta}(\lambda) \cong (L_{\mathfrak{g}}^{\beta\alpha}(\lambda))^{\triangle}$, then by Lemma 2.2 the result holds for $L_{\mathfrak{g}}^{\alpha\beta}(\lambda)$ as well. The case $A_1 \oplus A_1$ can be handled by constructing the corresponding representation. ∎

The main results of [Mc] and [Alv] were closely related to Theorem 5.3 for the cases of G_2 and C_2, respectively. For these rank two simple Lie algebras $\mathfrak{g}$, the lattices $L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$ were obtained by taking natural partial orders on the corresponding $\mathfrak{g}$-semistandard tableaux of section 4, and case analysis arguments were used to show that the mapping ϕ preserves weights and that the $\mathfrak{g}$-structure condition is satisfied. However, $\mathfrak{g}$-semistandard posets did not arise in their approach. If one is willing to depend entirely upon [Lit], then in this manner one can obtain Propositions 4.2 and 4.7 from Proposition 5.2 and Littelmann's analogue to Theorem 5.3 without using Lemma 3.1. But this approach would take at least as much (related) work and would not be as uniformly stated.

A_2 Admissible 1-block T	Weight $wt_{Lit}(T)$	Corresponding A_2-semistandard tableau
1	ω_α	1
2	$-\omega_\alpha + \omega_\beta$	2
3	$-\omega_\beta$	3
1 2	ω_β	1 2
1 3	$\omega_\alpha - \omega_\beta$	1 3
2 3	$-\omega_\alpha$	2 3

C_2 Admissible 2-block T	Weight $wt_{Lit}(T)$	Corresponding C_2-semistandard tableau
1 1	ω_α	1
2 2	$-\omega_\alpha + \omega_\beta$	2
3 3	$\omega_\alpha - \omega_\beta$	3
4 4	$-\omega_\alpha$	4
1 1 2 2	ω_β	1 2
1 1 3 3	$2\omega_\alpha - \omega_\beta$	1 3
1 2 3 4	$0\omega_\alpha + 0\omega_\beta$	2 3
2 2 4 4	$-2\omega_\alpha + \omega_\beta$	2 4
3 3 4 4	$-\omega_\beta$	3 4

G_2 Admissible 6-block T	Weight $wt_{Lit}(T)$	Corresponding G_2-semistandard tableau
1 1 1 1 1 1	ω_α	1
2 2 2 2 2 2	$-\omega_\alpha + \omega_\beta$	2
3 3 3 3 3 3	$2\omega_\alpha - \omega_\beta$	3
3 3 3 4 4 4	$0\omega_\alpha + 0\omega_\beta$	4
4 4 4 4 4 4	$-2\omega_\alpha + \omega_\beta$	5
5 5 5 5 5 5	$\omega_\alpha - \omega_\beta$	6
6 6 6 6 6 6	$-\omega_\alpha$	7

FIG. 5.2. *Admissible k-column blocks for Littelmann tableau, their weights, and their corresponding* $\mathfrak{g}$*-semistandard columns.*

G_2 Admissible 6-block T	Weight $wt_{Lit}(T)$	Corresponding G_2-semistandard tableau
1 1 1 1 1 1 2 2 2 2 2 2	ω_β	1 2
1 1 1 1 1 1 3 3 3 3 3 3	$3\omega_\alpha - \omega_\beta$	1 3
1 1 1 1 2 2 3 3 3 3 4 4	ω_α	1 4
1 1 2 2 2 2 3 3 4 4 4 4	$-\omega_\alpha + \omega_\beta$	1 5
2 2 2 2 2 2 4 4 4 4 4 4	$-3\omega_\alpha + 2\omega_\beta$	2 5
1 1 2 3 3 3 3 3 4 5 5 5	$2\omega_\alpha - \omega_\beta$	1 6
2 2 2 3 3 3 4 4 4 5 5 5	$0\omega_\alpha + 0\omega_\beta$	2 6
1 1 2 3 4 4 3 3 4 5 6 6	$0\omega_\alpha + 0\omega_\beta$	1 7
3 3 3 3 3 3 5 5 5 5 5 5	$3\omega_\alpha - 2\omega_\beta$	3 6
2 2 2 3 4 4 4 4 4 5 6 6	$-2\omega_\alpha + \omega_\beta$	2 7
3 3 3 3 4 4 5 5 5 5 6 6	$\omega_\alpha - \omega_\beta$	3 7
3 3 4 4 4 4 5 5 6 6 6 6	$-\omega_\alpha$	4 7
4 4 4 4 4 4 6 6 6 6 6 6	$-3\omega_\alpha + \omega_\beta$	5 7
5 5 5 5 5 5 6 6 6 6 6 6	$-\omega_\beta$	6 7

FIG. 5.2 (*continued*). *Admissible k-column blocks for Littelmann tableaux, their weights, and their corresponding* $\mathfrak{g}$*-semistandard columns.*

Our final result presents the $\mathfrak{g}$-semistandard lattices as answers to Stanley's Problem 3 [Sta1].

COROLLARY 5.4. *Let $\mathfrak{g}$ be a simple Lie algebra of rank two. Let $\lambda = (a, b)$, with $a, b \geq 0$. Then the $\mathfrak{g}$-semistandard lattices $L_{\mathfrak{g}}^{\beta\alpha}(\lambda)$ and $L_{\mathfrak{g}}^{\alpha\beta}(\lambda)$ are rank symmetric and rank unimodal. Moreover, the rank generating functions for these lattices are*

$$RGF_{A_2}(\lambda, q) = \frac{(1-q^{a+1})(1-q^{b+1})(1-q^{a+b+2})}{(1-q)(1-q)(1-q^2)},$$

$$RGF_{C_2}(\lambda, q) = \frac{(1-q^{a+1})(1-q^{b+1})(1-q^{a+b+2})(1-q^{a+2b+3})}{(1-q)(1-q)(1-q^2)(1-q^3)},$$

$$RGF_{G_2}(\lambda, q) = \frac{(1-q^{a+1})(1-q^{b+1})(1-q^{a+b+2})(1-q^{a+2b+3})(1-q^{a+3b+4})(1-q^{2a+3b+5})}{(1-q)(1-q)(1-q^2)(1-q^3)(1-q^4)(1-q^5)}.$$

In each case $|L_{\mathfrak{g}}^{\beta\alpha}(\lambda)| = |L_{\mathfrak{g}}^{\alpha\beta}(\lambda)|$, and these counts may be found by letting $q \to 1$.

Proof. In light of Theorem 5.3, apply Proposition 2.4. We have specialized the right-hand side quotient there using the data from Figure 5.1. □

6. Remarks. Stanley's Exercises 4.25 and 3.27 on Gaussian and pleasant posets have attracted some attention [Sta2]. A poset P with p elements is *Gaussian* if there exist positive integers $h_1, \ldots, h_p$ such that for all $m \geq 0$, the rank generating function of the lattice $J(P \times \mathbf{m})$ is $\Pi_{i=1}^{p}(1-q^{m+h_i})/(1-q^{h_i})$. In [Pr1], the sixth author and Stanley gave a uniform proof of the Gaussian property for all known Gaussian posets. That proof used an analogue of Theorem 5.3; it was based upon Seshadri's standard monomial basis theorem for the irreducible representations $X_n(m\omega_k)$, where the representations $X_n(\omega_k)$ are "minuscule." Now let P be our G_2-fundamental poset $P_{G_2}(0,1)$ of Figure 4.2. Please use Figure 3.2 to help visualize the G_2-semistandard poset $P_{G_2}^{\beta\alpha}(0,m)$ for $m \geq 0$. Note that $P_{G_2}^{\beta\alpha}(0,m)$ consists of $P \times \mathbf{m}$ together with some additional order relations. By Corollary 5.4, the rank generating function for $L_{G_2}^{\beta\alpha}(0,m) = J_{color}(P_{G_2}^{\beta\alpha}(0,m))$ is

$$\frac{(1-q^{m+1})(1-q^{m+2})(1-q^{2m+3})(1-q^{3m+4})(1-q^{3m+5})}{(1-q^1)(1-q^2)(1-q^3)(1-q^4)(1-q^5)}.$$

One could introduce a more general notion of "quasi-Gaussian" for a poset P by requiring that the elements of $P \times \mathbf{m}$ remain distinct when some additional (if any) order relations are introduced, and by allowing a more general product form for the generating function identity. Then the fundamental posets $P_{C_2}(0,1)$ and $P_{G_2}(1,0)$ of Figure 4.2 would also be quasi-Gaussian, but not Gaussian. In [DW] more will be said about the order relations added to $P \times \mathbf{m}$ above and the juxtaposition rules for the fundamental posets shown in Figures 3.2 and 3.3. For now, we note that these added order relations are similar to those added in the following example: The Catalan poset P_3 of Figure 1.1 can be obtained by adding order relations to the Gaussian poset $\mathbf{3} \times \mathbf{3}$; this corresponds to the restriction of $\mathfrak{sl}_6(\omega_3)$ to $\mathfrak{sp}_6(\omega_3)$.

Here is the C_2 example promised in the middle of section 4: The C_2-semistandard poset $P_{C_2}^{\beta\alpha}(1,1)$ is displayed in Figure 6.1. Also displayed is the corresponding C_2-semistandard lattice $L_{C_2}^{\beta\alpha}(1,1)$, with vertices labeled by the C_2-semistandard tableaux of shape $(1,1)$. The lattice $L_{C_2}^{\beta\alpha}(1,1)$ shown in Figure 6.1 looks similar in structure to the edge-colored lattice L displayed in Figure 6.2. In fact, this $L = J_{color}(P)$ for the two-color grid poset P displayed in Figure 6.2. Moreover, $P = Q_1 \triangleleft Q_2$ with $Q_1 \cong P_1$ and $Q_2 \cong P_2$ for the indecomposable two-color grid posets P_1 and P_2

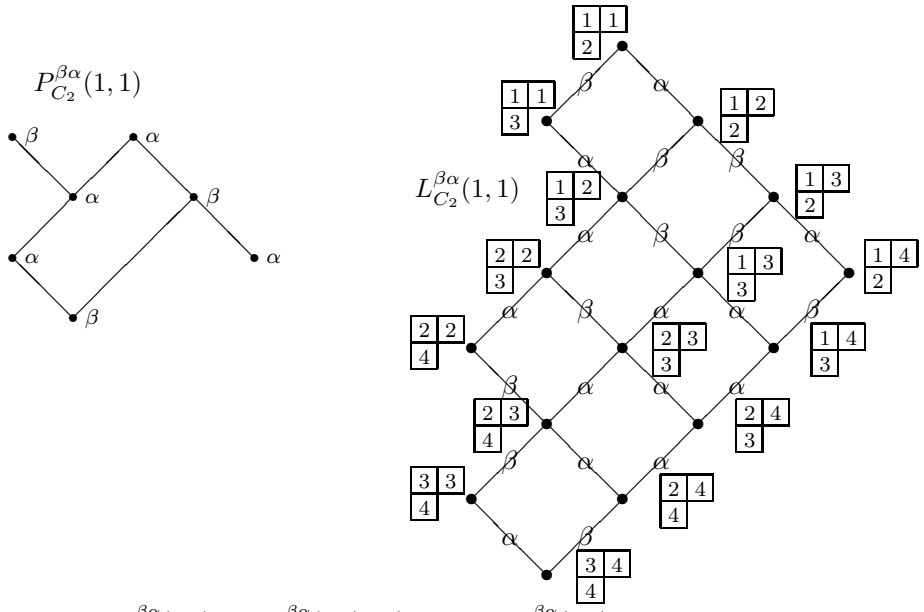


FIG. 6.1. $P^{\beta\alpha}_{C_2}(1,1)$ *and* $L^{\beta\alpha}_{C_2}(1,1)$. *(Vertices of* $L^{\beta\alpha}_{C_2}(1,1)$ *are indexed by* C_2*-semistandard tableaux.)*

displayed in Figure 6.2. And P_1 and P_2 look similar in structure to the fundamental $\mathfrak{g}$-semistandard posets presented in Figure 4.2. But it can be seen that L does not satisfy the structure condition for any 2×2 matrix M. Therefore, L cannot be a splitting poset for a representation, and so there is no hope of applying Proposition 2.4 to L. But L does have a "symmetric chain decomposition," and hence it is rank symmetric, rank unimodal, and "strongly Sperner."

It is possible to prove that the $\mathfrak{g}$-semistandard lattices, $\mathfrak{g} \in \{A_1 \oplus A_1, A_2, C_2, G_2\}$, are the only lattices of the kind we have been considering which can have the M-structure property for any 2×2 integer matrix M.

THEOREM 6.1 (see [Don2]). *Let P be a two-color grid poset which has the max property. If $L = J_{color}(P)$ has the M-structure property for some 2×2 integer matrix M, then L is a $\mathfrak{g}$-semistandard lattice, $\mathfrak{g} \in \{A_1 \oplus A_1, A_2, C_2, G_2\}$.*

THEOREM 6.2 (see [Don2]). *Let P be an indecomposable two-color grid poset. If $L = J_{color}(P)$ has the M-structure property for some 2×2 integer matrix M, then L is a $\mathfrak{g}$-fundamental lattice, $\mathfrak{g} \in \{A_1 \oplus A_1, A_2, C_2, G_2\}$.*

These two statements are combinatorial Dynkin diagram classification theorems: No Lie theory or algebraic concepts of any kind appear in their hypotheses, but the short list of Dynkin diagram-indexed rank two Cartan matrices plays the central role in their conclusions.

To apply Corollary 5.4 via Theorem 5.3, Proposition 5.2 was required: the elements of the $\mathfrak{g}$-semistandard lattices were matched up with tableaux of Littelmann. (But it is possible to directly obtain the total count results mentioned at the end of Corollary 5.4 with elementary combinatorial reasoning [ADLP].) The precise match-up required here should make one pessimistic about obtaining rank generating function identities similar to Corollary 5.4 for lattices $L = J_{color}(P)$ for general two-color grid posets P. This pessimism is intuitively heightened by the classification results above, which emphasize how special the $\mathfrak{g}$-semistandard lattices are. After representations for the cases listed in the introduction to this paper are constructed, Corollary 5.3 of [ADLP] notes that the $\mathfrak{g}$-semistandard lattices in those cases are strongly Sperner.

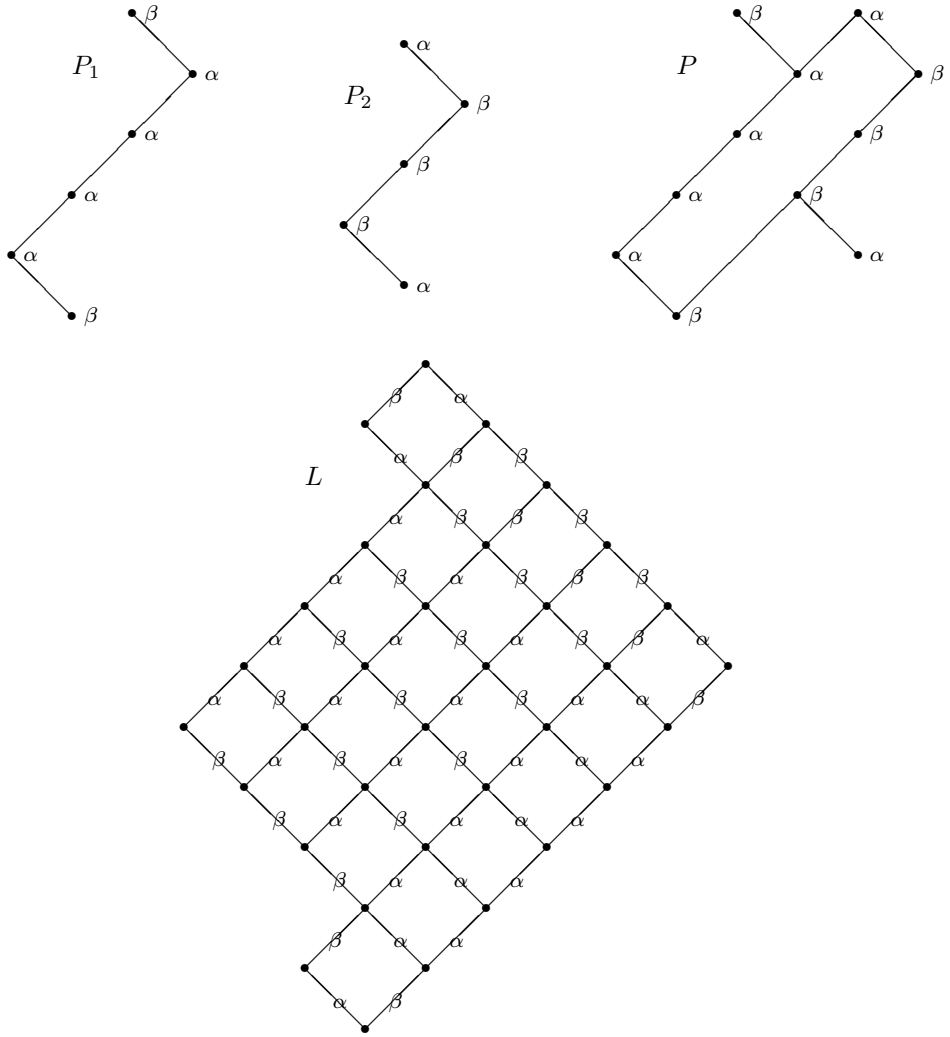


FIG. 6.2. *We can write* $P = Q_1 \triangleleft Q_2$ *with* $Q_i \cong P_i$ *(*$i = 1, 2$*). Below,* $L = J_{color}(P)$.

Although this approach cannot be used for the rest of the $\mathfrak{g}$-semistandard lattices, it is natural to hope that those lattices have this property. When addressing these extremal set theory issues, here it would now seem reasonable to attempt a combinatorial approach: Can one find symmetric chain decompositions of $L = J_{color}(P)$ for certain two-color grid posets P?

Although the rank two cases in Lie theory are much simpler than the general rank cases, it is also true in Lie theory that the key aspect of a higher rank case often reduces to consideration of that aspect for just the rank two cases. Various aspects of this rank two paper will be used for many higher rank cases in [DW]. The forms of the $\mathfrak{g}$-semistandard tableaux of section 4 may seem unmotivated to readers who are familiar only with [Hum]. Space and time permitting, much motivation could be supplied. Strict columns of length two arise because the second fundamental representation in each simple case can be realized as the "big piece" of the second exterior power of

the first fundamental representation. Standard monomial theory (and earlier papers concerning algebras with straightening laws) explain how the restricted concatenation of columns corresponds to the multiplication of "Plücker coordinates" for flag manifolds. Going further, it may be possible to "explain" the simple root colorings of the elements of the posets P in the spirit of the heaps of Stembridge, along the lines of Theorem 11.1 of [Pr1].

Our main result states that the $\mathfrak{g}$-semistandard lattices are splitting posets for their representations. For any representation, the crystal graph (of Kashiwara) is a splitting poset (cf. Lemma 3.6 of [Don1]). More generally, this is true for Stembridge's overarching crystal graph-like "admissible systems" [Stem]. The second author has observed that any admissible system for a given representation is "edge minimal" within the set of splitting posets for the representation: It contains no splitting poset for the representation as a proper subgraph. For all irreducible representations of types A_2 and C_2, it can be seen from [KN] that Kashiwara's crystal graphs are subgraphs of the corresponding $\mathfrak{g}$-semistandard lattices. The first, second, third, and fifth authors have recently shown that all $\mathfrak{g}$-semistandard lattices give rise to admissible systems. By replacing the step in section 5 of matching lattice elements with Littelmann's tableaux, this approach yields another proof Theorem 5.3. In [DW] we will consider most simple Lie algebras of arbitrary rank and uniformly define $\mathfrak{g}$-fundamental posets for their fundamental weights which have the following property: the longest element in the associated Bruhat order is "fully commutative." This definition is type-independent. Using these fundamental posets, as in section 4 we build $\mathfrak{g}$-semistandard posets and lattices for many representations. Along with this paper, this should start a new program: Find modular lattice splitting posets for all irreducible representations of all semisimple Lie algebras and show that they give rise to admissible systems. If these hopes are realized, these modular lattices (including the $\mathfrak{g}$-semistandard lattices) would in general contain "extra" edges with respect to the admissible system. But the lattices might be more combinatorially interesting than most or all admissible systems' directed graphs, and hopefully more accessible. One consequence might be the formulation of analogues of the Littlewood–Richardson tensor product rule in terms of manipulations of the underlying $\mathfrak{g}$-semistandard posets (or their analogues in the modular/nondistributive cases).

REFERENCES

[Alv] L. W. Alverson II, *Distributive Lattices and Representations of the Rank Two Simple Lie Algebras*, Master's thesis, Murray State University, Murray, KY, 2003.

[ADLP] L. W. Alverson II, R. G. Donnelly, S. J. Lewis, and R. Pervine, *Constructions of representations of rank two semisimple Lie algebras with distributive lattices*, Electron. J. Combin., 13 (2006), R109 (44 pp).

[Don1] R. G. Donnelly, *Extremal properties of bases for representations of semisimple Lie algebras*, J. Algebraic Combin., 17 (2003), pp. 255–282.

[Don2] R. G. Donnelly, *Dynkin diagram classification results satisfying certain structural properties*, in preparation.

[DLP] R. G. Donnelly, S. J. Lewis, and R. Pervine, *Constructions of representations of* $\mathfrak{o}(2n+1,\mathbb{C})$ *that imply Molev and Reiner–Stanton lattices are strongly Sperner*, Discrete Math., 263 (2003), pp. 61–79.

[DW] R. G. Donnelly and N. J. Wildberger, *Distributive lattice models for certain families of irreducible semisimple Lie algebra representations*, in preparation.

[Hum] J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory*, Springer-Verlag, New York, 1972.

[KN] M. Kashiwara and T. Nakashima, *Crystal graphs for representations of the q-analogue of classical Lie algebras*, J. Algebra, 165 (1994), pp. 295–345.

[Lit] P. Littelmann, *A generalization of the Littlewood–Richardson rule*, J. Algebra, 130 (1990), pp. 328–368.

[Mc] M. McClard, *Picturing Representations of Simple Lie Algebras of Rank Two*, Master's thesis, Murray State University, Murray, KY, 2000.

[Pr1] R. A. Proctor, *Bruhat lattices, plane partition generating functions, and minuscule representations*, European J. Combin., 5 (1984), pp. 331–350.

[Pr2] R. A. Proctor, *Solution of a Sperner conjecture of Stanley with a construction of Gelfand*, J. Combin. Theory Ser. A, 54 (1990), pp. 225–234.

[Sta1] R. P. Stanley, *Unimodal sequences arising from Lie algebras*, in Young Day Proceedings, T. V. Narayana et al., eds., Marcel Dekker, New York, 1980, pp. 127–136.

[Sta2] R. P. Stanley, *Enumerative Combinatorics, Vol.* 1, Wadsworth and Brooks/Cole, Monterey, CA, 1986.

[Sta3] R. P. Stanley, *Enumerative Combinatorics, Vol.* 2, Cambridge University Press, Cambridge, UK, 1999.

[Stem] J. Stembridge, *Combinatorial models for Weyl characters*, Adv. Math., 168 (2002), pp. 96–131.

SIAM J. DISCRETE MATH.
Vol. 23, No. 2, pp. 561–570

n-ARY QUASIGROUPS OF ORDER 4*

DENIS S. KROTOV† AND VLADIMIR N. POTAPOV†

Abstract. We characterize the set of all n-ary quasigroups of order 4: every n-ary quasigroup of order 4 is permutably reducible or semilinear. Permutable reducibility means that an n-ary quasigroup can be represented as a composition of k-ary and $(n-k+1)$-ary quasigroups for some k from 2 to $n-1$, where the order of arguments in the representation can differ from the original order. The set of semilinear n-ary quasigroups has a characterization in terms of Boolean functions.

Key words. Latin hypercube, n-ary quasigroup, reducibility

AMS subject classifications. 05B15, 20N05, 20N15, 94B25

DOI. 10.1137/070697331

1. Introduction. An algebraic system consisting of a finite set Σ of cardinality $|\Sigma| = q$ and an n-ary operation $f : \Sigma^n \to \Sigma$ uniquely invertible in each place is called an n-ary quasigroup of order q. The function f can also be referred to as an n-ary quasigroup of order q or, for short, an n-quasigroup. The value array of an n-quasigroup of order q is known as a Latin n-cube of order q (if $n = 2$, a Latin square). Furthermore, there is a one-to-one correspondence between the n-quasigroups and the distance 2 MDS codes in Σ^{n+1}.

It is known that for every n there exist exactly two equivalent n-quasigroups of order 2 and $3 \cdot 2^n$ n-quasigroups of order 3, which constitute one isotopy class (see, e.g., [LM98]). So, 4 is the first order for which a rich class of n-quasigroups exists. On the other hand, this order is of special interest for different areas of mathematics close to information theory. For example,

- the class of 1-perfect codes in $\{0,1\}^n$ of rank at most $n - \log_2(n+1) + 2$ (the minimum rank is $n - \log_2(n+1)$ for 1-perfect codes) is characterized in terms of n-quasigroups of order 4; see [AHS04] (so, our work completes this characterization);
- order 4 is the first order that is applicable for use in quasigroup stream ciphers; and
- from n-quasigroups of order 4, n-quasigroups of other orders can be constructed, giving examples of n-quasigroups with nontrivial properties (see, e. g., [Kro08a]).

In this paper, we show that every n-quasigroup of order 4 is permutably reducible or semilinear. Permutable reducibility means that the n-quasigroup can be represented as a repetition-free composition of quasigroups of smaller arities where the ordering of the arguments in the representation can differ from the original (see Definition 2.4). Semilinearity (Definition 2.5) means that the n-quasigroup can be obtained as a direct product of two n-quasigroups of order 2 modified by a Boolean function $\{0,1\}^n \to \{0,1\}$ (sometimes this construction is referred to as the wreath product

*Received by the editors July 16, 2007; accepted for publication (in revised form) October 6, 2008; published electronically February 6, 2009.
http://www.siam.org/journals/sidma/23-2/69733.html
†Sobolev Institute of Mathematics, prosp. Akademika Koptyuga, 4, Novosibirsk, 630090, Russia (krotov@math.nsc.ru, vpotapov@math.nsc.ru). The research of the first author was supported in part by the Russian Foundation for Basic Research, grant 08-01-00673. The research of the second author was supported in part by the Russian Foundation for Basic Research, grant 08-01-00671.

construction, but we should remember that this does not agree with the concept of the wreath product of groups).

In section 2 we introduce main concepts and notation. In section 3 we formulate the result (Theorem 3.1) and divide the proof into four subcases, Lemmas 3.2–3.5. Lemmas 3.4 and 3.5 are proved in sections 5 and 4, while Lemmas 3.2 and 3.3 follow from previous papers.

2. Main definitions.

DEFINITION 2.1. *An n-ary operation $Q : \Sigma^n \to \Sigma$, where Σ is a nonempty set, is called an* n-ary quasigroup *or* n-quasigroup (of order $|\Sigma|$) *if in the equality $z_0 = Q(z_1, \dots, z_n)$ knowledge of any n elements of $z_0, z_1, \dots, z_n$ uniquely specifies the remaining element* [Bel72].

The definition is symmetric with respect to the variables $z_0, z_1, \dots, z_n$, and sometimes it is convenient to use a symmetric form for the relation $z_0 = Q(z_1, \dots, z_n)$. For this reason, we will denote by $Q\langle z_0, z_1, \dots, z_n\rangle$ the corresponding predicate, i.e., the characteristic function of this relation. (In coding theory, the set corresponding to this predicate is known as a distance 2 MDS code.)

Given $\bar{y} = (y_1, \dots, y_n)$, we denote

$$\bar{y}^{[i]}[x] \triangleq (y_1, \dots, y_{i-1}, x, y_{i+1}, \dots, y_n);$$

similarly, we define $\bar{y}^{[i_1,i_2,\dots,i_k]}[x_{i_1}, x_{i_2}, \dots, x_{i_k}]$.

DEFINITION 2.2. *If we assign some fixed values to $l \in \{1, \dots, n\}$ variables in the predicate $Q\langle z_0, \dots, z_n\rangle$, then the $(n-l+1)$-ary predicate obtained corresponds to an $(n-l)$-quasigroup. Such a quasigroup is called a* retract *or* $(n-l)$-retract *of Q. If z_0 is not fixed, the retract is* principal.

DEFINITION 2.3. *By an* isotopy *we shall mean a collection of $n+1$ permutations $\tau_i : \Sigma \to \Sigma$, $i \in \{0, 1, \dots, n\}$. n-quasigroups f and g are called* isotopic *if for some isotopy $\bar{\tau} = (\tau_0, \tau_1, \dots, \tau_n)$ we have $f(x_1, \dots, x_n) \equiv \tau_0^{-1} g(\tau_1 x_1, \dots, \tau_n x_n)$, i.e., $f\langle x_0, x_1, \dots, x_n\rangle \equiv g\langle \tau_0 x_0, \tau_1 x_1, \dots, \tau_n x_n\rangle$.*

DEFINITION 2.4. *An n-quasigroup f is termed* permutably reducible *(in* [PK06]*, the term "decomposable" was used) if there exist $m \in \{2, \dots, n-1\}$, an $(n-m+1)$-quasigroup h, an m-quasigroup g, and a permutation $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ such that*

$$f(x_1, \dots, x_n) \equiv h(g(x_{\sigma(1)}, \dots, x_{\sigma(m)}), x_{\sigma(m+1)}, \dots, x_{\sigma(n)})$$

(i.e., f is a composition *of h and g). For short, we will omit the word "permutably" (with the exception of the main statements). If an n-quasigroup is not reducible, then it is* irreducible*. (In particular, all 2-quasigroups are irreducible.)*

DEFINITION 2.5. *We say that an n-quasigroup $f : \{0, 1, 2, 3\}^n \to \{0, 1, 2, 3\}$ is* standardly semilinear *if*

$$f\langle \bar{x}\rangle \leqq L\langle \bar{x}\rangle,$$

where

$$L\langle x_0, \dots, x_n\rangle \triangleq l(x_0) \oplus \dots \oplus l(x_n) \oplus 1, \quad l(0) = l(1) = 0, \quad l(2) = l(3) = 1$$

($\leqq$ means "$\leq$ everywhere"; $\oplus$ means "modulo-2 addition"); see, e.g., Figure 1*. An n-quasigroup of order* 4 *is called* semilinear *if it is isotopic to some standardly semilinear n-quasigroup.*

$f(x,y)$:

0	1	2	3
1	0	3	2
3	2	0	1
2	3	1	0

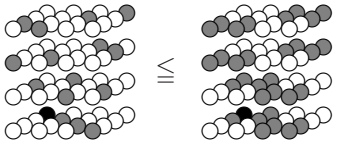
$f\langle z,x,y\rangle$: 

FIG. 1. *A standardly semilinear* 2*-quasigroup and the function* $L\langle\cdot\rangle$.

The set of standardly semilinear n-quasigroups has a simple characterization.

PROPOSITION 2.6. *The following relation is a bijection between the standardly semilinear* n*-quasigroups* f *and the Boolean functions* $\lambda : \{0,1\}^n \to \{0,1\}$*:*

$$f\langle x_0, x_1, \dots, x_n\rangle \equiv L\langle x_0, x_1, \dots, x_n\rangle \cdot \big(x_0 \oplus x_1 \oplus \dots \oplus x_1 \oplus \lambda(l(x_1), \dots, l(x_n))\big). \tag{1}$$

Proof. Consider a standardly semilinear n-quasigroup. Consider a set H consisting of 2^{n+2} points of $\{0,1,2,3\}^{n+1}$ with fixed values $l(x_1), \dots, l(x_n)$.

The number of 1's of $f\langle\cdot\rangle$ in H is 2^n (indeed, by the definition of an n-quasigroup, every n-tuple $x_1, \dots, x_n$ corresponds to exactly one 1). Moreover, since f is standardly semilinear, all these 1's belong to $H_1 \triangleq (\bar{x} \in H \mid L\langle\bar{x}\rangle = 1)$. Since there are no two 1's that differ in only one coordinate, all these 1's simultaneously have either an even or an odd coordinate sum. In the even case, define $\lambda(l(x_1), \dots, l(x_n)) = 1$; in the odd case, $\lambda(l(x_1), \dots, l(x_n)) = 0$. Then (1) is automatically true. □

So, the number of the standardly semilinear n-quasigroups is 2^{2^n}. Multiplying by the number $3^{n+1}2$ of different functions isotopic to L, we obtain an approximate number of the semilinear n-quasigroups. The exact number is $3^{n+1}2^{2^n+1} - 8 \cdot 6^n$ [PK06, Theorem 1], where $-8 \cdot 6^n$ is explained by the fact that affine Boolean functions (and only affine, i.e., of type $\lambda(z_1, \dots, z_n) = b_0 \oplus b_1 z_1 \oplus \dots \oplus b_n z_n$, $b_i \in \{0,1\}$) correspond to n-quasigroups majorized by more than one isotope of L.

In the rest of the paper, unless otherwise stated, we consider only order 4 n-quasigroups over $\Sigma = \{0,1,2,3\}$.

3. Main result. The main result is the following theorem.

THEOREM 3.1. *Every* n*-quasigroup of order* 4 *is permutably reducible or semilinear.*

The basic characteristic of an n-quasigroup f, which divides our proof into four subcases, is the maximum arity of its irreducible retract. Denote this value by $\kappa(f)$; then, $2 \le \kappa(f) < n$. The line of reasoning in the proof of Theorem 3.1 is inductive, so we can assume that the irreducible retracts are semilinear.

LEMMA 3.2 (case $\kappa = n-1$ [PK06, Lemma 4]). *If an* n*-quasigroup* f *of order* 4 *has a semilinear* $(n-1)$*-retract, then it is permutably reducible or semilinear.*

LEMMA 3.3 (case $2 < \kappa \le n-3$ [Kro08b]). *Let* f *be an* n*-quasigroup of arbitrary order and* $\kappa(f) \in \{3, \dots, n-3\}$*. Then* f *is permutably reducible.*

In [Kro08a], an example of an irreducible n-quasigroup of order 4 whose $(n-1)$-retracts are all reducible is constructed for every even $n \ge 4$. So, the assumption of Lemma 3.3 cannot be extended to the case $\kappa = n-2$. Nevertheless, in section 5 we will prove the following.

LEMMA 3.4 (case $\kappa = n-2$). *Let* $n \ge 5$*. If an* n*-quasigroup* f *of order* 4 *has a semilinear permutably irreducible* $(n-2)$*-retract and all the* $(n-1)$*-retracts are permutably reducible, then* f *is permutably reducible or semilinear.*

The last case, announced in [Pot06], will be proved in section 4.

LEMMA 3.5 (case $\kappa = 2$). *Let $n \geq 5$; let f be an n-quasigroup of order 4, and let all its k-retracts with $2 < k < n$ be permutably reducible. Then f is permutably reducible.*

Proof of Theorem 3.1. The validity of the theorem for $n \leq 4$ (and even for $n \leq 5$) is proved by exhaustion. Assume, by induction, that all m-quasigroups of order 4 with $m < n$ are reducible or semilinear. Consider an n-quasigroup f of order 4. It has an irreducible $\kappa(f)$-retract, which is semilinear by inductive assumption. Depending on the value of $\kappa(f) = 2, 3 \ldots n{-}3, n{-}2, n{-}1$, the statement of the theorem follows from one of Lemmas 3.5, 3.3, 3.4, and 3.2. ∎

4. Proof of Lemma 3.5. We will prove a stronger variant (Lemma 4.2) of the statement. It uses the following concept.

DEFINITION 4.1. *An n-quasigroup f is called* completely reducible *if it is permutably reducible and all its principal retracts of arity more than* 2 *are permutably reducible (equivalently, f can be represented as a composition of $n-1$ binary quasigroups; e.g., $f(x_1, x_2, x_3, x_4, x_5) = f_1(f_2(x_1, x_3), f_3(f_4(x_2, x_5), x_4))$).*

LEMMA 4.2. *Let all the principal* 3- *and* 4*-retracts of an n-quasigroup f of order* 4 *($n \geq 5$) be permutably reducible. Then f is completely reducible.*

DEFINITION 4.3. *An n-quasigroup f is called* normalized *if for all $i \in \{1, \ldots, n\}$ and $a \in \Sigma$ it is true that $f(\bar{0}^{[i]}[a]) = a$.*

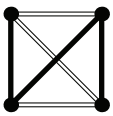
Denote by Γ the set of all (four) normalized binary quasigroups of order 4. It is straightforward that the operations from Γ are associative and commutative (they are isomorphic to the additive groups Z_2^2 and Z_4), and we will use the form $a \star b$ instead of $\star(a, b)$ to write the result of $\star \in \Gamma$.

Let $K_n = \langle V(K_n), E(K_n) \rangle$ be the complete graph with n vertices associated with the arguments $x_1, \ldots, x_n$ of an n-ary operation. For the edges, we will use the short notation like $x_i x_j$. For any normalized n-quasigroup f we define the edge coloring $\mu_f : E(K_n) \to \Gamma$ in the following way: the color $\mu_f(x_i x_j)$ of an edge $x_i x_j \in E(K_n)$ is defined as the binary operation $\star$ such that $f(\bar{0}^{[i,j]}[x_i, x_j]) \equiv x_i \star x_j$.

PROPOSITION 4.4. *Let f be an n-quasigroup, $n \geq 5$, and let all* 3- *and* 4*-retracts of f be reducible. Then the coloring μ_f of K_n satisfies the following:*

(A) *Every triangle is colored by at most two colors.*

(B) *If a tetrahedron is colored by two colors with three edges of each color, then it includes a one-color triangle; i.e., the following fragment is forbidden:*

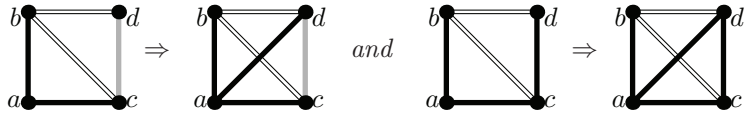


Proof. Every 3-retract of f is a composition of two binary operations which yields (A).

Consider the 4-retract f_4 of f that corresponds to some four vertices (the other variables are fixed by 0). Since it is reducible and normalized, it can be represented as a composition of some normalized 2-quasigroup and 3-quasigroup. The 3-quasigroup is a 3-retract of f and, in its turn, can be represented as a composition of two normalized 2-quasigroups. So, $f_4(x, y, z, u)$, up to permutation of arguments, has the form $(x \star y) \circ (u \diamond v)$ or $x \circ (y \star (u \diamond v))$ for some $\star, \circ, \diamond \in \Gamma$. As follows from the hypothesis of (B), two of these three operations coincide. Thus, there are only four types of decomposition of f_4: $(x \star y) \circ (u \star v)$, $x \star (y \circ u \circ v)$, $x \star y \star (u \circ v)$, $x \circ (y \star (u \circ v))$. In any case, (B) holds. ∎

PROPOSITION 4.5. *Assume that an edge coloring μ of K_n satisfies* (A) *and* (B).

Then for each pairwise different $a, b, c, d \in V(K_n)$ the condition $\mu(ab) = \mu(ac) \neq \mu(bc) = \mu(bd) \neq \mu(cd)$ implies $\mu(ad) = \mu(ab)$. That is,



Proof. Obviously, any other variant for $\mu(ad)$ contradicts (A) or (B). □

The following proposition is easy to check.

PROPOSITION 4.6. *Let f and g be reducible 3-quasigroups of order 4, and let $f(x, y, 0) \equiv g(x, y, 0)$, $f(x, 0, z) \equiv g(x, 0, z)$, and $f(0, y, z) \equiv g(0, y, z)$. Then $f \equiv g$.*

Remark. Indeed, Proposition 4.6 holds for every order with the extra condition that f is a composition of two different or associative 2-quasigroups. The similar statement for n-quasigroups with $n > 3$ holds for an arbitrary order without extra conditions [KPS08, Theorem 1]: if two reducible n-quasigroups coincide on every n-tuple with one zero, then they are identical.

COROLLARY 4.7. *Let f and g be n-quasigroups of order 4 ($n \geq 3$) whose principal 3-retracts are all reducible. Assume that $f(\bar{0}^{[i,j]}[y, z]) \equiv g(\bar{0}^{[i,j]}[y, z])$ for every $i, j \in \{1, \dots, n\}$ and $y, z \in \Sigma$. Then $f \equiv g$.*

Proof. The equality $f(\bar{x}) \equiv g(\bar{x})$ is proved by induction on the number of nonzero elements in $\bar{x}$, using the reducibility of 3-retracts and Proposition 4.6.

For example, to prove that $f(1, 2, 3, 2, 1, \bar{0}) = g(1, 2, 3, 2, 1, \bar{0})$ we can consider the 3-retracts $f^3(x, y, z) \triangleq f(1, 2, x, y, z, \bar{0})$ and $g^3(x, y, z) \triangleq g(1, 2, x, y, z, \bar{0})$. By the induction assumption f^3 and g^3 meet the hypothesis of Proposition 4.6. Thus, $f^3 \equiv g^3$, and, in particular, $f^3(3, 2, 1) = g^3(3, 2, 1)$. □

The following proposition is the key statement in the proof of Lemma 4.2.

PROPOSITION 4.8. *Assume that an edge coloring $\mu : E(K_n) \to \Gamma$ of the graph K_n meets* (A) *and* (B). *Then there exists a completely reducible n-quasigroup f such that $\mu_f = \mu$.*

Before proving Proposition 4.8 by induction, we consider one auxiliary statement, which will be used in the induction step. We say that an edge $xy \in E(K_n)$ is *inner* with respect to some edge coloring μ of K_n if for any $z \in V(K_n) \setminus \{x, y\}$ it is true that $\mu(xz) = \mu(yz)$.

PROPOSITION 4.9. *Assume that an edge coloring μ of K_n meets* (A) *and* (B). *Then K_n contains an inner edge.*

Proof. Consider an arbitrary sequence of edges $e_1, e_2, \dots, e_k$ that satisfies the following:

(C) for every $j \in \{1, \dots, k-1\}$ the edges e_j and e_{j+1} are adjacent and $\mu(e_j) = \mu(e_j \triangle e_{j+1}) \neq \mu(e_{j+1})$ (where $\triangle$ means the symmetrical difference between two sets).

Denote by a_j the element from $e_j \setminus e_{j+1}$.

Claim (*). We claim that for every i, j, $1 \leq i < j \leq k$, and $d \in e_j$ the vertices a_i and d are different and $\mu(a_i d) = \mu(e_i)$ (see Figure 2). We will show this by induction on $j - i$.

If $j - i = 1$, the claim follows from (C). If $j - i = 2$, the claim follows from Proposition 4.5 ($a := a_i$, $b := a_{i+1}$). Assume $j - i > 2$. By the inductive assumption,

$$\begin{aligned} &\mu(e_i) = \mu(a_i a_{i+1}) = \mu(a_i a_{i+2}) \\ \neq\ &\mu(e_{i+1}) = \mu(a_{i+1} a_{i+2}) = \mu(a_{i+1} d) \\ \neq\ &\mu(e_{i+2}) = \mu(a_{i+2} d). \end{aligned}$$

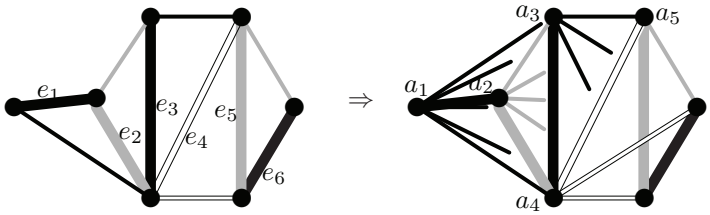


FIG. 2. *An example of a sequence $e_1, \dots, e_k$ from the proof of Proposition* 4.8.

Consequently, $a_i \neq d$, and, by Proposition 4.5 ($a := a_i$, $b := a_{i+1}$, $c := a_{i+2}$), we have $\mu(a_i d) = \mu(e_i)$. Claim (*) is proved.

So, all $a_1, \dots, a_{k-1}$ are mutually different, and thus there exists a maximum sequence $e_1, e_2, \dots, e_k$ satisfying (C). Then, its maximality and (*) imply that the edge e_k is inner. ∎

Proof of Proposition 4.8. We will proceed by induction on n. If $n = 3$, then the statement is trivially true. Assume that Proposition 4.8 holds for $(n-1)$-quasigroups. Consider an inner edge $e \in E(K_n)$. Without loss of generality we can assume $e = x_{n-1}x_n$, $\mu(e) = \star$. Denote by μ^{n-1} the restriction of the coloring μ on $K_{n-1} \subset K_n$. By the inductive assumption, there exists a completely reducible $(n-1)$-quasigroup g such that $\mu_g = \mu^{n-1}$. Then $f(x_1, \dots, x_n) \triangleq g(x_1, x_2, \dots, x_{n-2}, x_{n-1} \star x_n)$ is a desired n-quasigroup (indeed, $\mu_f(x_{n-1}x_n) = \star = \mu(x_{n-1}x_n)$; if $i < j < n$, then $\mu_f(x_i x_j) = \mu_g(x_i x_j) = \mu(x_i x_j)$; if $i < n-1$, then $\mu_f(x_i x_n) = \mu_g(x_i x_{n-1}) = \mu(x_i x_{n-1}) = \mu(x_i x_n)$, where the last equality follows from the innerness of $x_{n-1}x_n$). ∎

Proof of Lemma 4.2. Let f be an n-quasigroup of order 4 whose 3- and 4-retracts are all reducible. Without loss of generality we assume that f is normalized (otherwise, we can normalize it, applying an appropriate isotopy). Then, by Proposition 4.4, the corresponding edge coloring μ_f of the graph K_n satisfies (A) and (B). By Proposition 4.8, there exists a completely reducible n-quasigroup g with $\mu_g \equiv \mu_f$. By Corollary 4.7, f and g are identical. ∎

5. Proof of Lemma 3.4. In the proof, we will use the following three propositions. The first simple one, on a representation of a reducible n-quasigroup with an irreducible $(n-1)$-retract, holds for an arbitrary order.

PROPOSITION 5.1. *Assume that a reducible n-quasigroup D ($n \geq 3$) of an arbitrary order has an irreducible $(n-1)$-retract $F\langle x_0, \dots, x_{n-1}\rangle \equiv D\langle x_0, \dots, x_{n-1}, 0\rangle$. Then there are $i \in \{0, \dots, n\}$ and a 2-quasigroup h such that $h(x, 0) \equiv x$ and*

$$D\langle x_0, \dots, x_n\rangle \equiv F\langle x_0, \dots, x_{i-1}, h(x_i, x_n), x_{i+1}, \dots, x_{n-1}\rangle. \tag{2}$$

Proof. Since D is reducible, $D\langle x_0, \dots, x_n\rangle$ can be represented as $H\langle f(\bar{x}'), \bar{x}''\rangle$, where $\bar{x}'$ and $\bar{x}''$ are disjoint groups of variables, each containing at least two variables. If x_n is grouped with more than one other variable, then fixing x_n gives a reducible retract, which contradicts the irreducibility of F. So, we conclude that for some $i \in \{0, \dots, n-1\}$ there exists one of the following two representations of D:

$$\begin{aligned} D\langle x_0, \dots, x_n\rangle &\equiv G\langle g(x_i, x_n), \tilde{x}\rangle, \\ D\langle x_0, \dots, x_n\rangle &\equiv g\langle G(\tilde{x}), x_i, x_n\rangle, \end{aligned} \tag{3}$$

where $\tilde{x} \triangleq (x_0, \dots, x_{i-1}, x_{i+1}, \dots, x_{n-1})$ and G and g are $(n-1)$- and 2-quasigroups. Moreover, the existence of a representation of the first type implies the existence of a representation of the second type, and vice versa:

$$g\langle G(\tilde{x}), x_i, x_n\rangle \equiv \begin{cases} 1 & \text{if } G(\tilde{x}) = g(x_i, x_n) \\ 0 & \text{if } G(\tilde{x}) \neq g(x_i, x_n) \end{cases} \equiv G\langle g(x_i, x_n), \tilde{x}\rangle.$$

So, we can assume that (3) holds. Put $\gamma(x_i) \triangleq g(x_i, 0)$. Then, $F\langle x_0, \dots, x_{n-1}\rangle \equiv G\langle\gamma(x_i), \tilde{x}\rangle$, and (2) holds with $h(x_i, x_n) \triangleq \gamma^{-1} g(x_i, x_n)$. □

In what follows, permutations $\sigma : \Sigma \to \Sigma$ will be denoted by the value lists $(\sigma(0), \sigma(1), \sigma(2), \sigma(3))$; denote $Id \triangleq (0,1,2,3)$.

PROPOSITION 5.2 (on autotopies of a semilinear n-quasigroup). *Assume f is a standardly semilinear n-quasigroup. Denote by π the permutation* (1,0,3,2)*. Then for every different $i, j \in \{0, \dots, n\}$ the following hold:*

(a) $f\langle\bar{x}\rangle \equiv f\langle\bar{x}^{[i,j]}[\pi x_i, \pi x_j]\rangle$*, where* $\bar{x} \triangleq (x_0, \dots, x_n)$*;*

(b) *if* $f\langle\bar{x}\rangle \equiv f\langle\bar{x}^{[i,j]}[\mu x_i, \nu x_j]\rangle$ *holds for some other pair of nonidentity permutations* $(\mu, \nu) \neq (\pi, \pi)$ *and* $n \geq 3$*, then f is reducible.*

Proof. (a) It is straightforward that $f\langle\bar{x}^{[i]}[\pi x_i]\rangle \equiv L\langle\bar{x}\rangle - f\langle\bar{x}\rangle$, where $L\langle\cdot\rangle$ is from Definition 2.5. So, $f\langle\bar{x}^{[i,j]}[\pi x_i, \pi x_j]\rangle \equiv L\langle\bar{x}\rangle - (L\langle\bar{x}\rangle - f\langle\bar{x}\rangle) \equiv f\langle\bar{x}\rangle$.

(b) Without loss of generality assume that $i = 1$, $j = 2$. Put

$$\begin{aligned} \alpha(x, y) &\triangleq f(x, y, \bar{0}), \\ \beta(x, \bar{z}) &\triangleq f(x, 0, \bar{z}), \qquad \bar{z} \triangleq (z_1, \dots, z_{n-2}), \\ \gamma(x) &\triangleq f(x, 0, \bar{0}). \end{aligned}$$

Assume there exists a pair (μ, ν) that satisfies the hypothesis of (b). Then $\alpha(x, y) \equiv \alpha(\mu x, \nu y)$. It is easy to see that the permutation ν does not have fixed points. So, ν is either a cyclic permutation or an involution ((2,3,0,1) or (3,2,1,0)) different from $\pi = (1,0,3,2)$. In any case, we can derive the following.

Claim (*). For each $v \in \Sigma$ there exist permutations $\rho_v, \tau_v : \Sigma \to \Sigma$ such that $f(x, y, \bar{z}) \equiv f(\rho_v x, \tau_v y, \bar{z})$ and $\tau_v v = 0$ (in other words, the group of permutations τ admitting $f(x, y, \bar{z}) \equiv f(\rho x, \tau y, \bar{z})$ for some ρ acts transitively on Σ, i.e., has only one orbit).

Case 1. If ν is a cyclic permutation, then v, νv, $\nu^2 v$, $\nu^3 v$ are pairwise different; so, one of the pairs (Id, Id), (μ, ν), (μ^2, ν^2), (μ^3, ν^3) can be chosen as (ρ_v, τ_v), proving (*).

Case 2. If ν is (2,3,0,1) or (3,2,1,0), then v, νv, πv, $\nu\pi v$ are pairwise different, and (ρ_v, τ_v) can be chosen from (Id, Id), (μ, ν), (π, π), $(\mu\pi, \nu\pi)$.

Claim (*) is proved. Then,

$$\begin{aligned} f(x, y, \bar{z}) &\equiv f(\rho_y x, \tau_y y, \bar{z}) \equiv f(\rho_y x, 0, \bar{z}) \equiv \beta(\rho_y x, \bar{z}) \\ &\equiv \beta(\gamma^{-1}\alpha(\rho_y x, 0), \bar{z}) \equiv \beta(\gamma^{-1}\alpha(\rho_y x, \tau_y y), \bar{z}) \equiv \beta(\gamma^{-1}\alpha(x, y), \bar{z}), \end{aligned}$$

and thus f is reducible provided $n \geq 3$. □

The next proposition concerns 2-quasigroups of order 4, and the proof is straightforward.

PROPOSITION 5.3. *Let s and t be 2-quasigroups. Denote $s_i(x) \triangleq s(x, i)$ and $t_i(x) \triangleq t(x, i)$. Let $s_0 = t_0 = Id$, and let for every i either $t_i s_i^{-1} = Id$ or $t_i s_i^{-1} =$* (1,0,3,2)*. Then either $s \equiv t$ or for some permutation ϕ the 2-quasigroup $s'(x, y) \triangleq s(x, \phi y)$ is standardly semilinear.*

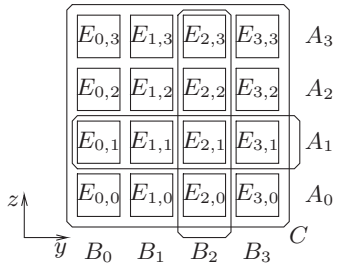
Proof. Denote $\pi \triangleq (1,0,3,2)$; observe that $\pi = \pi^{-1}$. Assume that $s \not\equiv t$. Then there are at least two different elements $i, j \in \{1,2,3\}$ for which $t_i s_i^{-1} = \pi$. Denote by k the third element, i.e., $\{i,j,k\} = \{1,2,3\}$. The permutation t_i has no fixed points; otherwise there is a contradiction with t_0; similarly, $s_i = \pi t_i$ has no fixed points. So, t_i and, similarly, t_j belong to $\{(2,3,0,1), (2,3,1,0), (3,2,0,1), (3,2,1,0)\}$. The only variant for t_k is (1,0,3,2). Then, $s'(x,y) \triangleq s(x,\phi y)$ is standardly semilinear with $\phi \triangleq (0,k,i,j)$. ∎

Proof of Lemma 3.4. Assume C is an n-quasigroup of order 4. Assume all the $(n-1)$-retracts of C are reducible and C has a semilinear irreducible $(n-2)$-retract E. Without loss of generality assume that

$$E\langle x_0,\dots,x_{n-2}\rangle \equiv C\langle x_0,\dots,x_{n-2},0,0\rangle$$

and E is standardly semilinear. We will use the following notation for retracts of C (the table illustrates their mutual arrangement, where the last and second-to-last coordinates of Σ^{n+1} are thought of as ordinate and abscissa, respectively; for example, B_2 corresponds to fixing the abscissa by (2):

$$\begin{aligned} E_{a,b}\langle x_0,\dots,x_{n-2}\rangle &\triangleq C\langle x_0,\dots,x_{n-2},a,b\rangle,\\ A_b\langle x_0,\dots,x_{n-2},y\rangle &\triangleq C\langle x_0,\dots,x_{n-2},y,b\rangle,\\ B_a\langle x_0,\dots,x_{n-2},z\rangle &\triangleq C\langle x_0,\dots,x_{n-2},a,z\rangle. \end{aligned}$$



Since A_0 is reducible and fixing $y := 0$ leads to the irreducible E, by Proposition 5.1 we have

$$A_0\langle x_0,\dots,x_{n-2},y\rangle \equiv E\langle x_0,\dots,x_{i-1},h(x_i,y),x_{i+1},\dots,x_{n-2}\rangle \tag{4}$$

for some $i \in \{0,\dots,n-2\}$ and 2-quasigroup h such that $h(x_i,0) \equiv x_i$.

From (4), we see that all the retracts $E_{a,0}$, $a \in \Sigma$, are isotopic to E. Similarly, we can get the following.

Claim (*). All the retracts $E_{a,b}$, $a,b \in \Sigma$ are isotopic to E.

Then, we conclude that a representation similar to (4) is valid for every $b \in \Sigma$:

$$A_b\langle x_0,\dots,x_{n-2},y\rangle \equiv E_{0,b}\langle x_0,\dots,x_{i_b-1},h_b(x_{i_b},y),x_{i_b+1},\dots,x_{n-2}\rangle$$

for some $i \in \{0,\dots,n-2\}$ and 2-quasigroup h_b such that $h_b(x,0) \equiv x$.

Claim (**). We claim that i_b does not depend on b. Indeed, assume, for example, that $i_1 = 0$ and $i_2 = 1$, i.e.,

$$\begin{aligned} A_1\langle x_0,\dots,x_{n-2},y\rangle &\equiv E_{0,1}\langle h_1(x_0,y),x_1,x_2,\dots,x_{n-2}\rangle,\\ A_2\langle x_0,\dots,x_{n-2},y\rangle &\equiv E_{0,2}\langle x_0,h_2(x_1,y),x_2,\dots,x_{n-2}\rangle. \end{aligned}$$

Then, fixing x_0 in the first case leads to a retract isotopic to E; fixing x_0 in the second case leads to a reducible retract (recall that $n \geq 5$). But, analogously to (*), these two retracts are isotopic; this contradicts the irreducibility of E and proves (**).

Without loss of generality we can assume that $i_b = 0$, i.e.,

$$A_b\langle x_0,x_1,\tilde{x}_2,y\rangle \equiv E_{0,b}\langle h_b(x_0,y),x_1,\tilde{x}_2\rangle; \tag{5}$$

here and later $\tilde{x}_2 \triangleq (x_2, \dots, x_{n-2})$. Similarly, we can assume without loss of generality that either

$$B_a\langle x_0, x_1, \tilde{x}_2, z\rangle \equiv E_{a,0}\langle g_a(x_0, z), x_1, \tilde{x}_2\rangle \tag{6}$$

or

$$B_a\langle x_0, x_1, \tilde{x}_2, z\rangle \equiv E_{a,0}\langle x_0, g_a(x_1, z), \tilde{x}_2\rangle, \tag{7}$$

where 2-quasigroups g_a satisfy $g_a(x, 0) \equiv x$.

Using (5) and (6), we derive

$$\begin{aligned} C\langle x_0, x_1, \tilde{x}_2, y, z\rangle &\equiv A_z\langle x_0, x_1, \tilde{x}_2, y\rangle \\ &\equiv E_{0,z}\langle h_z(x_0, y), x_1, \tilde{x}_2\rangle \\ &\equiv B_0\langle h_z(x_0, y), x_1, \tilde{x}_2, z\rangle \\ &\equiv E_{0,0}\langle g_0(h_z(x_0, y), z), x_1, \tilde{x}_2\rangle, \end{aligned} \tag{8}$$

which means that C is reducible, because $f(x, y, z) \triangleq g_0(h_z(x, y), z)$ must be a 3-quasigroup. So, it remains to consider the case (7). Consider two subcases.

Case 1. The 2-quasigroup g_a does not depend on a; denote $g \triangleq g_a$. Then, repeating the first three steps of (8) and applying (7), we derive that

$$C\langle x_0, x_1, \tilde{x}_2, y, z\rangle \equiv E_{0,0}\langle h_0(x_0, y), g(x_1, z), \tilde{x}_2\rangle,$$

and C is reducible.

Case 2. For some fixed a we have $g_0 \neq g_a$; denote $s_i(x) \triangleq g_0(x, i)$, $t_i(x) \triangleq g_a(x, i)$, and $r_i(x) \triangleq h_i(x, a)$. From (5), we see that

$$E_{a,0}\langle x_0, x_1, \tilde{x}_2\rangle \equiv E_{0,0}\langle r_0(x_0), x_1, \tilde{x}_2\rangle, \tag{9}$$

$$E_{a,b}\langle x_0, x_1, \tilde{x}_2\rangle \equiv E_{0,b}\langle r_b(x_0), x_1, \tilde{x}_2\rangle. \tag{10}$$

From (7), we see that

$$E_{0,b}\langle x_0, x_1, \tilde{x}_2\rangle \equiv E_{0,0}\langle x_0, s_b(x_1), \tilde{x}_2\rangle, \tag{11}$$

$$E_{a,b}\langle x_0, x_1, \tilde{x}_2\rangle \equiv E_{a,0}\langle x_0, t_b(x_1), \tilde{x}_2\rangle. \tag{12}$$

Applying consecutively (11), (10), (12), and (9), we find that for each b the retract $E = E_{0,0}$ satisfies

$$E\langle x_0, x_1, \tilde{x}_2\rangle \equiv E_{0,b}\langle \dots\rangle \equiv E_{a,b}\langle \dots\rangle \equiv E_{a,0}\langle \dots\rangle \equiv E\langle r_0 r_b^{-1} x_0, t_b s_b^{-1} x_1, \tilde{x}_2\rangle.$$

By Proposition 5.2, the irreducibility of E means that $t_b s_b^{-1} \in \{Id, (1,0,3,2)\}$ for every b. By Proposition 5.3, for some permutation ϕ the 2-quasigroup $s(x, z) \triangleq g_0(x, \phi z)$ is standardly semilinear. Since a composition of standardly semilinear quasigroups is a standardly semilinear quasigroup, we see that B_0 is a semilinear $(n-1)$-quasigroup. Lemma 3.2 completes the proof. ∎

Acknowledgments. The authors thank the referees for their work in reviewing the manuscript and the audience of the seminar "Coding Theory" in the Sobolev Institute of Mathematics for their patience during the reporting of this result.

REFERENCES

[AHS04] S. V. Avgustinovich, O. Heden, and F. I. Solov'eva, *The classification of some perfect codes*, Des. Codes Cryptogr., 31 (2004), pp. 313–318.

[Bel72] V. D. Belousov, *n-Ary Quasigroups*, Shtiintsa, Kishinev, 1972 (in Russian).

[Kro08a] D. S. Krotov, *On irreducible n-ary quasigroups with reducible retracts*, European J. Combin., 29 (2008), pp. 507–513.

[Kro08b] D. S. Krotov, *On reducibility of n-ary quasigroups*, Discrete Math., 308 (2008), pp. 5289–5297.

[KPS08] D. S. Krotov, V. N. Potapov, and P. V. Sokolova, *On reconstructing reducible n-ary quasigroups and switching subquasigroups*, Quasigroups Related Systems, 16 (2008), pp. 55–67.

[LM98] C. F. Laywine and G. L. Mullen, *Discrete Mathematics Using Latin Squares*, Wiley, New York, 1998.

[Pot06] V. N. Potapov, *On completely commutatively reducible n-quasigroups*, in Proceedings of the 16th International School-Seminar on Synthesis and Complexity of Controlling Systems, St. Petersburg, Russia, 2006, pp. 88–91 (in Russian).

[PK06] V. N. Potapov and D. S. Krotov, *Asymptotics for the number of n-quasigroups of order* 4, Siberian Math. J., 47 (2006), pp. 720–731.

SIAM J. DISCRETE MATH.
Vol. 23, No. 2, pp. 571–595

PARTIAL GRÖBNER BASES FOR MULTIOBJECTIVE INTEGER LINEAR OPTIMIZATION*

VÍCTOR BLANCO† AND JUSTO PUERTO†

Abstract. This paper presents a new methodology for solving multiobjective integer linear programs (MOILP) using tools from algebraic geometry. We introduce the concept of partial Gröbner basis for a family of multiobjective programs where the right-hand side varies. This new structure extends the notion of Gröbner basis for the single objective case to the case of multiple objectives, i.e., when there is a partial ordering instead of a total ordering over the feasible vectors. The main property of these bases is that the partial reduction of the integer elements in the kernel of the constraint matrix by the different blocks of the basis is zero. This property allows us to prove that this new construction is a test family for a family of multiobjective programs. An algorithm "á la Buchberger" is developed to compute partial Gröbner bases, and two different approaches are derived, using this methodology, for computing the entire set of Pareto-optimal solutions of any MOILP problem. Some examples illustrate the application of the algorithm, and computational experiments are reported on several families of problems.

Key words. multiple objective optimization, integer programming, Gröbner bases, test sets

AMS subject classifications. 90C29, 90C10, 13P10

DOI. 10.1137/070698051

1. Introduction.

The multiobjective paradigm appeared in economic theory in the nineteenth century in the seminal works by Edgeworth [14] and Pareto [30] to define an economic equilibrium. Mathematically, the multiobjective optimization approach consists of determining the maximal (minimal) elements of a partially ordered set. This problem was already addressed by Cantor [7], Cayley [8], and Hausdorff [21] at the end of the nineteenth century. Since then, multiobjective programming (including multicriteria optimization) has been a fruitful research field within the areas of applied mathematics, operations research, and economic theory. Excellent textbooks and survey papers are available in the literature; the interested reader is referred to the books by Sawaragi, Nakayama, and Tanino [32], Chankong and Haimes [9], Yu [45], Miettinen [28], or Ehrgott, Figueira, and Gandibleux [19], and to the surveys in [17, 18].

The importance of multiobjective optimization is not due only to its theoretical implications but also to its many applications. Witnesses of that are the large number of real-world decision problems that appear in the literature formulated as multiobjective programs. These include flowshop scheduling [24], analysis in finance [17], railway network infrastructure capacity [13], vehicle routing problems [25, 34], or trajectory optimization [36], among many others.

Multiobjective programs are formulated as optimization (without lost of generality, we restrict ourselves to the minimization case) problems over feasible regions with at least two objective functions. Usually, it is not possible to minimize all of the objective functions simultaneously, since the objective functions induce a partial

*Received by the editors July 23, 2007; accepted for publication (in revised form) October 17, 2008; published electronically February 6, 2009. This research was partially supported by Ministerio de Educación y Ciencia under grant MTM2007-67433-C02-01.
http://www.siam.org/journals/sidma/23-2/69805.html
†Departmento de Estadística e Investigación Operative, Universidad de Sevilla, 41012 Sevilla, Spain (vblanco@us.es, puerto@us.es).

order over the vectors in the feasible region, so a different notion of solution is needed. A feasible vector is said to be Pareto-optimal (or nondominated) if no other feasible vector has componentwise smaller objective values, with at least one strict inequality.

This paper studies multiobjective integer linear programs (MOILP). Thus, we assume that all objective functions and constraints that define the feasible region are linear and that the feasible vectors have nonnegative integer components.

There are nowadays several exact methods to solve MOILP problems. Two of them, multiobjective implicit enumeration [46, 47] and multiobjective dynamic programming [26], claim to be of general use and have attracted the attention of researchers in the last several years. Nevertheless, although, in principle, they may be applied to any number of objectives, one can mainly find, in the literature, applications to biobjective problems. Moreover, some other methods even do not provide the entire set of Pareto-optimal solutions, but provide the supported ones (those that can be obtained as solutions of linearly scalarized programs).

On the other hand, there are several methods that apply to biobjective problems but that do not extend to the general case. Thus, one can see that there are two thresholds in multiobjective programming: a first step from one to two objectives, and a second, and deeper one, from two to more than two objectives.

In terms of complexity, it is worth noting that most MOILP problems are NP-hard [16]. Even when the single-objective problem is polynomially solvable, the multiobjective version may become NP-hard. This is the case of spanning tree [20] and minimum-cost flow problems [15], among others. Therefore, computational efficiency is not an issue when analyzing MOILP. The important point is to develop tools that can handle these problems and that give insights into their intrinsic nature. The goal of this paper is to present a new general methodology for solving MOILP using tools borrowed from algebraic geometry. The usage of algebraic geometry tools in integer programming (single criterion) is not new (see [10], [22], [41], [23], [44], [43]). The main idea is to compute a Gröbner basis for certain toric ideals (related to the constraints matrix) with a monomial order induced by the objective function.

Gröbner bases were introduced by Buchberger in 1965 in his Ph.D. thesis [6]. He named them Gröbner bases paying tribute to his advisor Wolfgang Gröbner. This theory emerged as a generalization, from the one variable case to the multivariate polynomial case, of the greatest common divisor. One of the outcomes of Gröbner bases theory was its application to integer programming, first published by Conti and Traverso [10]. This paper opened a new research line, followed by a number of authors, consisting of the application of algebraic geometry tools for solving integer programs.

In [22], Hoşten and Sturmfels gave two ways to implement the Conti and Traverso algorithm that improve in some cases the branch–and-bound algorithm to exactly solve integer programs. Thomas presented in [41] a geometric point of view of the Buchberger algorithm as a method to obtain solutions of an integer program. Later, Thomas and Weismantel [43] improved the Buchberger algorithm in its application to solve integer programs introducing truncated Gröbner bases. At the same time, Urbaniak, Weismantel, and Ziegler [44] published a clear geometric interpretation of the reduction steps of this kind of algorithm in the original space (decision space). The interested reader can find excellent descriptions of this methodology in the books by Adams and Loustaunau [2], Sturmfels [37], Cox, Little, and O'Shea [12], or Bertsimas and Weismantel [5], and in the papers by Aardal, Weismantel, and Wolsey [1], Sturmfels [38, 39], Sturmfels and Thomas [40], and Thomas [42].

Our main contribution is to adapt some of the above-mentioned tools from algebraic geometry to solve MOILP problems. We present an algorithm to exactly solve multiobjective problems, i.e., providing the whole set of Pareto-optimal solutions (supported and nonsupported ones). One of the main advantages of our approach is that the number of objective functions does not significantly increase the difficulty. A new geometric approach of the notion of reduction based on a partial ordering is given. This reduction allows us to extend the concept of Gröbner basis when a partial ordering rather than a total ordering is considered over $\mathbb{N}^n$. We call these new structures partial Gröbner bases or p-Gröbner bases. We prove that p-Gröbner bases can be generated in a finite number of steps by a variation of the Buchberger algorithm. The main property of a p-Gröbner basis being that, for each pair in $\mathbb{Z}^n \times \mathbb{Z}^n_+$ with first component in $\mathrm{Ker}(A)$, the reduction by maximal chains in the basis is the zero set.

We propose two versions of the same algorithm to solve multiobjective integer programs based on this new construction. Our first approach consists of three stages. The first one uses only the constraint matrix of the problem, and it produces a system of generators for the toric ideal $\Im_A$ (or its geometric representation I_A). In the second step, a p-Gröbner basis is built using the initial basis given by the system of generators computed in the first step. This step requires us to fix the objective matrix, since it induces the partial order used in the reduction steps. Once the right-hand side vector is fixed, in the third step, the Pareto-optimal solutions are obtained. This computation uses the new concept of partial reduction of an initial feasible solution by the p-Gröbner basis.

This algorithm extends, to some extent, Hoşten and Sturmfels' algorithm [22] for integer programs because if we apply our method to single-objective problems, partial reductions and p-Gröbner bases coincide with the standard notions of reductions and Gröbner bases, respectively.

Our second approach is based on the original idea by Conti and Traverso [10]. It consists of using the big-M method that results in an increasing number of variables, in order to have an initial system of generators. Moreover, this approach also provides an initial feasible solution. Therefore, the first step in the former variant of the algorithm can be ignored, and the third step is highly simplified. In any case, our first version (the one extending the Hoşten and Sturmfels approach) has proved to be more efficient than this second one, since the computation of a p-Gröbner basis is highly sensitive to the number of variables.

Both algorithms have been implemented in MAPLE 10. We report on some computational experiments based on the first version of the algorithm and on two different families of problems with different number of objective functions.

The rest of the paper is organized as follows. In section 2 we give the notation, the formulation of the problem, and its algebraic codification. We also introduce here the notion of test family and its geometric description. Section 3 presents the definition of p-Gröbner basis, based on the notion of partial reduction. Here, we also state the relationship between test families and p-Gröbner bases: the reduced p-Gröbner basis for a family of multiobjective programs varying the right-hand side coincides with the minimal test family for that family. At the end of the section, an example illustrates all of the above concepts. Section 4 is devoted to the results of the computational experiments and its analysis. Here, we solve several families of MOILP, report on the performance of the algorithms, and draw some conclusions on its results and their implications.

2. The problem and its translation. The goal is to solve the MOILP in its standard form:

$$\begin{aligned} &\min\ (c_1\,x, \ldots, c_k\,x) \\ &\text{subject to (s.t.)} \sum_{j=1}^{n} a_{ij}\,x_j = b_i \quad i = 1, \ldots, m, \\ &\qquad x_j \in \mathbb{Z}_+, \quad j = 1, \ldots, n, \end{aligned} \tag{1}$$

with b_i nonnegative integers for $i = 1, \ldots, m$, $c_l \in \mathbb{Z}_+^n$ for $l = 1, \ldots, k$, $x = (x_1, \ldots, x_n)$, and the constraints define a polytope (bounded). For the sake of simplicity, at times, we use a vector notation and denote $A = (a_{ij}) \in \mathbb{Z}^{m\times n}$, $b = (b_i) \in \mathbb{Z}_+^m$, and $C = (c_{ij}) \in \mathbb{Z}_+^{k\times n}$. In the following, problem (1) is referred to as $MIP_{A,C}(b)$, and we denote by $MIP_{A,C}$ the family of multiobjective problems where the right-hand side varies.

The reader may note that there is no loss of generality in our approach to multiobjective integer linear programming, since any general MOILP problem with inequality constraints and rational components in A, b, and C can be transformed to a problem in the above standard form.

It is clear that the problem $MIP_{A,C}(b)$ is not a usual optimization problem since the objective function is a vector, thus inducing a partial order among its feasible solutions. Hence, solving the above problem requires an alternative concept of solution, namely, the set of nondominated or Pareto-optimal points (vectors).

A feasible vector $\widehat{x} \in \mathbb{R}^n$ is said to be a *Pareto-optimal* or *nondominated solution* of $MIP_{A,C}(b)$ if there is no other feasible vector y such that

$$c_j\,y \le c_j\,\widehat{x} \qquad \text{for all } j = 1, \ldots, k$$

with at least one strict inequality for some j.

If x is a Pareto-optimal solution, the vector $(c_1\,x, \ldots, c_k\,x) \in \mathbb{R}^k$ is called *efficient*.

We say that a feasible point y is *dominated* by a feasible point x if $c_i\,x \le c_i\,y$ for all $i = 1, \ldots, k$, with at least one strict inequality. According to the above concept, solving a multiobjective problem consists of finding its entire set of Pareto-optimal solutions, including those that have the same objective values.

From the objective function C, we obtain a partial order over $\mathbb{Z}^n$ as follows:

$$x \prec_C y :\Longleftrightarrow C\,x \lneqq C\,y \quad or \quad x = y,$$

where $Cx \lneqq Cy$ stands for $Cx \le Cy$ and $Cx \ne Cy$.

Observe that since $C \in \mathbb{Z}_+^{k\times n}$, the above relation is not complete. Hence, there may exist incomparable vectors (those x, $y \in \mathbb{Z}_+^n$ such that neither $x \prec_C y$ nor $y \prec_C x$). We use this partial order induced by the objective function of problem $MIP_{A,C}$ as the input for the multiobjective integer programming algorithm developed in this paper.

Remark 2.1. Note that distinct solutions with the same objective values are incomparable under $\prec_C$. This order can be refined so that those solutions with the same objective values are comparable. Consider the binary relation

$$x \preceq_C y :\Longleftrightarrow \begin{cases} C\,x \lneqq C\,y, & \text{or} \\ Cx = Cy \text{ and } x \prec_{lex} y. \end{cases}$$

This alternative order allows us to rank those solutions that have the same objective values using the lexicographical order of their components.

The above partial order $\preceq_C$ permits us to solve a simplified version of the multi-objective problem after introducing the following equivalence relation in $\mathbb{Z}^n$:

$$x \sim_C y :\Longleftrightarrow Cx = Cy.$$

In this version, we obtain solutions in $\mathbb{Z}^n/\sim_C$. The reader may note that when solving the problem with the order $\preceq_C$, one would obtain only a representative element of each class of Pareto-optimal solutions (the lexicographically smallest). With those efficient values $\{v_1, \ldots, v_t\}$, the remaining solutions can be obtained solving the following system of diophantine equations, in x, for each v_i, $i = 1, \ldots, t$:

$$\begin{cases} Cx & = v_i, \\ Ax & = b, \\ x & \in \mathbb{Z}^n_+. \end{cases}$$

Remark 2.2. In some cases, the order $\prec_C$ can be refined to be adapted to specific problems. This is the case when slack variables appear in mathematical programs. Two feasible solutions (x, s_1) and (x, s_2), where s_1 and s_2 are the slack components, have the same objective values. The order $\prec_C$ considers both solutions as incomparable, although they are the same because we are looking just for the x-component of the solution. In these cases, we consider the following refined partial order in $\mathbb{Z}^n \times \mathbb{Z}^r$:

$$(x, s) \prec^s_C (y, s') :\Longleftrightarrow \begin{cases} C\,x \lneqq C\,y, & \text{or} \\ Cx = Cy \text{ and } s \prec_{lex} s', \end{cases}$$

where $x, y \in \mathbb{Z}^n_+$ are the actual decision variables and $s, s' \in \mathbb{Z}^r_+$ are the slack variables of our problem.

In the following, we will use partial order $\prec_C$ unless it is explicitly specified.

Our matrix A is encoded in the set

$$J_A = \{\{u, v\} : u, v \in \mathbb{N}^n, u - v \in \mathrm{Ker}(A)\}. \tag{2}$$

Let $\pi : \mathbb{N}^n \longrightarrow \mathbb{Z}^n$ denote the map $x \mapsto Ax$. Given a right-hand side vector b in $\mathbb{Z}^n$, the set of feasible solutions to $MIP_{A,C}(b)$ constitutes $\pi^{-1}(b)$, the preimage of b under this map. In the rest of the paper, we identify the discrete set of points $\pi^{-1}(b)$ with its convex hull, and we call it the b-fiber of $MIP_{A,C}$. Thus, $\pi^{-1}(b)$ or the b-fiber of $MIP_{A,C}$ is the polyhedron defined by the convex hull of all feasible solutions to $MIP_{A,C}(b)$.

For any pair $\{u, v\}$, with $u, v \in \mathbb{N}^n$, we define the set $setlm(u, v)$ as follows:

$$setlm(u, v) = \begin{cases} \{u\} & \text{if } v \prec_C u, \\ \{v\} & \text{if } u \prec_C v, \\ \{u, v\} & \text{if } u \text{ and } v \text{ are incomparable by } \prec_C. \end{cases}$$

The reader may note that $setlm(u, v)$ is the set of degrees of the leading monomials according to identification $\{u, v\} \mapsto x^u - x^v \in \mathbb{R}[x_1, \ldots, x_n]$, induced by partial order $\prec_C$.

From the above definition, $setlm(u, v)$ may have more than one leading term, since $\prec_C$ is only a partial order. To account for all this information, we denote by $\mathcal{F}(u, v)$ the set of triplets

$$\mathcal{F}(u, v) = \{(u, v, w) : w \in setlm(u, v)\}.$$

The above concept extends to any finite set of pairs of vectors in $\mathbb{N}^n$, accordingly. For a pair of sets $\mathbf{u} = \{u_1, \ldots, u_t\}$ and $\mathbf{v} = \{v_1, \ldots, v_t\}$, the corresponding set of ordered pairs is

$$\mathcal{F}(\mathbf{u}, \mathbf{v}) = \{(u_i, v_i, w) : w \in setlm(u_i, v_i), i = 1, \ldots, t\}.$$

$\mathcal{F}(\mathbf{u}, \mathbf{v})$ can be partially ordered based on the third component of its elements. Therefore, we can see $\mathcal{F}(\mathbf{u}, \mathbf{v})$ as a directed graph $G(E, V)$, where V is identified with the elements of $\mathcal{F}(\mathbf{u}, \mathbf{v})$ and $((u_i, v_i, w'), (u_j, v_j, w)) \in E$ if $(u_i, v_i, w), (u_j, v_j, w') \in V$ and $w' \prec_C w$. We are interested in the maximal ordered chains of G. Note that they can be efficiently computed by different methods, e.g., [4], [33].

The above concepts are clarified in the following example.

Example 2.1. Let $\mathbf{u} = \{(2,3), (0,2), (3,0), (2,1), (1,1)\}$, $\mathbf{v} = \{(1,4), (1,3), (4,2), (1,2), (1,0)\}$, and $\prec_C$ be the partial order induced by the matrix

$$C = \begin{bmatrix} 2 & 1 \\ 3 & 5 \end{bmatrix},$$

then, $setlm((2,3),(1,4)) = \{(2,3),(1,4)\}$, $setlm((0,2),(1,3)) = \{(1,3)\}$, $setlm((3,0),(4,2)) = \{(4,2)\}$, $setlm((2,1),(1,2)) = \{(2,1),(1,2)\}$, and $setlm((1,1),(1,0)) = \{(1,1)\}$. Now, by definition, we have

$$\begin{aligned}\mathcal{F}(\mathbf{u}, \mathbf{v}) = \{&\big((2,3),(1,4),(2,3)\big), \quad \big((2,3),(1,4),(1,4)\big), \big((0,2),(1,3),(1,3)\big),\\ &\big((3,0),(4,2),(4,2)\big), \quad \big((2,1),(1,2),(2,1)\big), \big((2,1),(1,2),(1,2)\big),\\ &\big((1,1),(1,0),(1,1)\big)\}.\end{aligned}$$

Figure 1 corresponds to the directed graph associated with $\mathcal{F}(\mathbf{u}, \mathbf{v})$, according to the partial ordering induced by C. There are four maximal chains:
$M_1 = \{\big((3,0),(4,2),(4,2)\big), \big((2,3),(1,4),(2,3)\big), \big((0,2),(1,3),(1,3)\big), \big((2,1),(1,2),(2,1)\big), \big((1,1),(1,0),(1,1)\big)\}$,
$M_2 = \{\big((3,0),(4,2),(4,2)\big), \big((2,3),(1,4),(2,3)\big), \big((0,2),(1,3),(1,3)\big), \big((2,1),(1,2),(1,2)\big), \big((1,1),(1,0),(1,1)\big)\}$,
$M_3 = \{\big((2,3),(1,4),(1,4)\big), \big((0,2),(1,3),(1,3)\big), \big((2,1),(1,2),(2,1)\big), \big((1,1),(1,0),(1,1)\big)\}$,
$M_4 = \{\big((2,3),(1,4),(1,4)\big), \big((0,2),(1,3),(1,3)\big), \big((2,1),(1,2),(1,2)\big), \big((1,1),(1,0),(1,1)\big)\}$.

For any pair of sets $\mathbf{u} = \{u_1, \ldots, u_t\}$ and $\mathbf{v} = \{v_1, \ldots, v_t\}$, with $\{u_i, v_i\} \in J_A$, for all $i = 1, \ldots, t$, the corresponding set $\mathcal{F}(\mathbf{u}, \mathbf{v})$ may also be seen as a set of pairs in $\mathbb{Z}^n \times \mathbb{Z}^n_+$ through the following map:

$$\begin{aligned}\phi\colon \mathbb{N}^n \times \mathbb{N}^n \times \mathbb{N}^n &\longrightarrow \mathbb{Z}^n \times \mathbb{Z}^n_+\\ (u, v, w) &\mapsto (u - v, w).\end{aligned}$$

We denote by $I_A = \phi(\mathcal{F}(J_A))$, i.e.,

$$I_A = \{(u - v, w) : u - v \in \mathrm{Ker}(A), w = setlm(u, v)\}.$$

It is clear that the maximal chains $F_1, \ldots, F_r$ of the image of $\mathcal{F}(\mathbf{u}, \mathbf{v})$ under ϕ with respect to the order $\prec_C$ over the second components satisfy the following properties:

1. F_i is totally ordered by the second components with respect to $\prec_C$ for $i = 1, \ldots, r$.

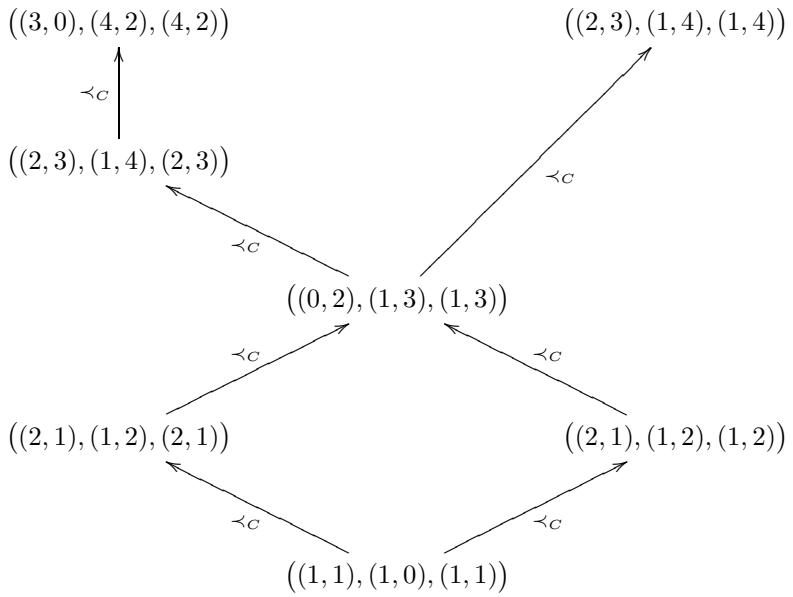


FIG. 1. *Hasse diagram of the graph associated with the data in Example* 2.1.

2. For all $(\alpha, \beta) \in F_i$, $i = 1, \ldots, r$, $A(\beta - \alpha) = A\beta$.

The map ϕ and the above properties allow us to define the notion of test family for $MIP_{A,C}$. This notion is analogous to the concept of test set for a family of single objective integer programs when we have a partial order rather than a total order over $\mathbb{N}^n$ [41]. Test families are instrumental for finding the Pareto-optimal set of each member $MIP_{A,C}(b)$ of the family of MOILP.

DEFINITION 2.1 (test family). *A finite collection* $\mathcal{G} = \{\mathcal{G}_C^1, \ldots, \mathcal{G}_C^r\}$ *of sets in* $\mathbb{Z}^n \times \mathbb{Z}_+^n$ *is a test family for* $MIP_{A,C}$ *if and only if*

(1) $\mathcal{G}_C^j$ *is totally ordered by the second component with respect to* $\prec_C$ *for* $j = 1, \ldots, r$.
(2) *For all* $(g, h) \in \mathcal{G}_C^j$, $j = 1, \ldots, r$, $A(h - g) = Ah$.
(3) *If* $x \in \mathbb{N}^n$ *is a dominated solution for* $MIP_{A,C}(b)$, *with* $b \in \mathbb{Z}_+^n$, *there is some* $\mathcal{G}_C^j$ *in the collection, and* $(g, h) \in \mathcal{G}_C^j$ *such that* $x - g \prec_C x$.
(4) *If* $x \in \mathbb{N}^n$ *is a Pareto-optimal solution for* $MIP_{A,C}(b)$, *with* $b \in \mathbb{Z}_+^n$, *then for all* $(g, h) \in \mathcal{G}_C^j$ *and for all* $j = 1, \ldots, r$, *either* $x - g$ *is infeasible or* $x - g$ *is incomparable to* x.

Given a test family for $MIP_{A,C}$, there is a natural approach for finding the entire Pareto-optimal set. Suppose we wish to solve $MIP_{A,C}(b)$ for which x^* is a feasible solution.

If x^* is dominated, then there is some j and $(g, h) \in \mathcal{G}_C^j$ such that $x^* - g$ is feasible and $x^* - g \prec_C x^*$, whereas, for the remaining chains, there may exist some (g, h) such that $x^* - g$ is feasible but incomparable to x^*. We keep track of all of them.

If x^* is nondominated, we have to keep it as an element in our current solution set. Then, reducing x^* by the chains in the test family, we can only obtain either incomparable feasible solutions, that we maintain in our structure, or infeasible solutions that are discarded.

The above two cases lead us to generate the following set. From x^*, we compute the set of incumbent solutions:

$$IS(x^*) := \{y^* : y^* = x^* - g_{j_i}, (g_{j_i}, h_{j_i}) \text{ is the largest element } (g,h) \text{ in the chain } \mathcal{G}_C^i \text{ such that } x^* - g \text{ is feasible }, i = 1, \dots, r\}.$$

Now, the scheme proceeds recursively on each element of the set $IS(x^*)$. Finiteness of the above scheme is clear, since we are generating a search tree with bounded depth (cardinality of the test family) and bounded width, each element in the tree has at most r (number of chains) followers. The correctness of this approach is ensured, since any pair of Pareto-optimal solutions must be connected by a reduction chain through elements in the test family (see Theorem 2.1 and Corollary 2.1).

The above approach assumes that a feasible solution to $MIP_{A,C}(b)$ is known (thus implying that the problem is feasible). Methods to detect infeasibility and to get an initial feasible solution are connected to solving diophantine systems of linear equations; the interested reader is referred to [31] for further details.

The following lemmas help us in describing the geometric structure of a test family for multiobjective integer linear problems.

LEMMA 2.1 (Gordan–Dickson lemma, Theorem 5 in [11]). *If $P \subseteq \mathbb{N}^n$, $P \neq \emptyset$, then there exists a minimal subset $\{p_1, \dots, p_m\} \subseteq P$ that is finite and unique such that $p \in P$ implies $p_j \leq p$ (componentwise) for at least one $j = 1, \dots, m$.*

LEMMA 2.2. *There exists a unique, minimal, finite set of vectors $\alpha_1, \dots, \alpha_k \in \mathbb{N}^n$ such that the set $\mathcal{L}_C$ of all dominated solutions in all fibers of $MIP_{A,C}$ is a subset of $\mathbb{N}^n$ of the form*

$$\mathcal{L}_C = \bigcup_{j=1}^{k} (\alpha_j + \mathbb{N}^n).$$

Proof. The set of dominated solutions of all problems $MIP_{A,C}$ is

$$\mathcal{L}_C = \{\alpha \in \mathbb{N}^n : \exists \beta \in \mathbb{N}^n, \text{ with } A\beta = A\alpha \text{ and } \beta \prec_C \alpha\}.$$

Let α be an element in $\mathcal{L}_C$ and β a Pareto-optimal point in the fiber $\pi^{-1}(A\alpha)$ that satisfies $\beta \prec_C \alpha$. Then, for any $\gamma \in \mathbb{N}^n$, $A(\alpha + \gamma) = A(\beta + \gamma)$, $\alpha + \gamma, \beta + \gamma \in \mathbb{N}^n$, and $\beta + \gamma \prec_C \alpha + \gamma$, because the cost matrix C has only nonnegative coefficients. Therefore, $\alpha + \gamma$ is a feasible solution dominated by $\beta + \gamma$ in the fiber $\pi^{-1}(A(\alpha + \gamma))$. Then, $\alpha + \gamma \in \mathcal{L}_C$ for all $\gamma \in \mathbb{N}^n$, so $\alpha + \mathbb{N}^n \subseteq \mathcal{L}_C$. By Lemma 2.1, we conclude that there exists a minimal set of elements $\alpha_1, \dots, \alpha_k \in \mathbb{N}^n$ such that $\mathcal{L}_C = \bigcup_{j=1}^{k} (\alpha_j + \mathbb{N}^n)$. ∎

Once elements $\alpha_1, \dots, \alpha_k$ generating $\mathcal{L}_C$ (in the sense of the above result) have been obtained, one can compute the maximal chains of the set $\{\alpha_1, \dots, \alpha_k\}$ with respect to the partial order $\prec_C$. We denote by $\mathcal{C}_C^1, \dots, \mathcal{C}_C^\mu$ these maximal chains and set $\mathcal{L}_C^i = \bigcup_{t=1}^{k_i} (\alpha_t^i + \mathbb{N}^n)$, where $\alpha_t^i \in \mathcal{C}_C^i$ for $t = 1, \dots, k_i$ and $i = 1, \dots, \mu$. For details about maximal chains, upper bounds on its cardinality and algorithms to compute them for a partially ordered set, the reader is referred to [4].

It is clear that, with this construction, we have $\mathcal{L}_C = \bigcup_{i=1}^{\mu} \mathcal{L}_C^i$.

Next, we describe a finite family of sets $\mathcal{G}_{\prec_C} \subseteq \text{Ker}(A) \cap \mathbb{Z}^n$ and prove that it is indeed a test family for $MIP_{A,C}$.

Let $\mathcal{G}_{\prec_C} = \{\mathcal{G}_{\prec_C}^i\}_{i=1}^{\mu}$, being

$$\mathcal{G}_{\prec_C}^i = \{(g_{ij}^k, h_{ij}^k) = (\alpha_j^i - \beta_{ij}^k, \alpha_j^i), j = 1, \dots k_i, k = 1, \dots, m_{ij}\}, i = 1, \dots, \mu \tag{3}$$

the maximal chains of $\mathcal{G}_{\prec_C}$ (with respect to the order $\prec_C$ over the second components) and where $\alpha_1^i, \ldots, \alpha_{k_i}^i$ are the unique minimal elements of $\mathcal{L}^i_{\prec_C}$ and $\beta_{ij}^1, \ldots, \beta_{ij}^{m_{ij}}$ are the Pareto-optimal solutions to the problem $MIP_{A,C}(A\alpha_j^i)$.

In the next section, we give an algorithm that explicitly constructs $\mathcal{G}_{\prec_C}$. Notice that for fixed i, j and k, $g_{ij}^k = (\alpha_j^i - \beta_{ij}^k)$ is a point in the subspace $S = \{x \in \mathbb{Q}^n : Ax = 0\}$, i.e., in the 0-fiber of $MIP_{A,C}$. Geometrically we think of $(\alpha_j^i - \beta_{ij}^k, \alpha_j^i)$ as the oriented vector $\overrightarrow{g}_{ij}^k = \overrightarrow{[\beta_{ij}^k, \alpha_j^i]}$ in the $A\alpha_j^i$-fiber of $MIP_{A,C}$. The vector is directed from the Pareto-optimal point β_{ij}^k to the nonoptimal point α_j^i due to the minimization criterion in $MIP_{A,C}$, which requires us to move away from expensive points. Subtracting the point $\overrightarrow{g}_{ij}^k = \alpha_j^i - \beta_{ij}^k$ from the feasible solution γ gives the new solution $\gamma - \alpha_j^i + \beta_{ij}^k$, which is equivalent to translating $\overrightarrow{g}_{ij}^k$ by a nonnegative integer vector.

Consider an arbitrary fiber of $MIP_{A,C}$ and a feasible lattice point γ in this fiber. For each vector $\overrightarrow{g}_{ij}^k$ in $\mathcal{G}_{\prec_C}$, check whether $\gamma - g_{ij}^k$ is in $\mathbb{N}^n$. At γ, draw all such possible translations of vectors from $\mathcal{G}_{\prec_C}$. The head of the translated vector is also incident at a feasible point in the same fiber as γ, since g_{ij}^k is in the 0-fiber of $MIP_{A,C}$. We do this construction for all feasible points in all fibers of $MIP_{A,C}$. From Lemma 2.2 and the definition of $\mathcal{G}_{\prec_C}$, it follows that no vector $(\alpha_j^i - \beta_{ij}^k, \alpha_j^i)$ in $\mathcal{G}_{\prec_C}$ can be translated by a ν in $\mathbb{N}^n$ such that its tail meets a Pareto-optimal solution on a fiber unless the obtained vector is incomparable to the Pareto-optimal point β_{ij}^k.

THEOREM 2.1. *The above construction builds a connected directed graph in every fiber of $MIP_{A,C}$. The nodes of the graph are all the lattice points in the fiber, and (γ, γ') is an edge of the directed graph if $\gamma' = \gamma - g_{ij}^k$ for some i, j, and k. Any directed path of this graph is nonincreasing with respect to the partial order $\prec_C$.*

Proof. Pick a fiber of $MIP_{A,C}$ and, at each feasible lattice point, construct all possible translations of the vector $\overrightarrow{g}_{ij}^k$ from the set $\mathcal{G}^i_{\prec_C}$ as described above. Let α be a lattice point in this fiber. By Lemma 2.2, $\alpha = \alpha_j^i + \nu$ for some $i \in \{1, \ldots, t\}$ and $\nu \in \mathbb{Z}_+^n$. Now, since the point α_k' defined as $\alpha_k' = \beta_{ij}^k + \nu$ also lies in the same fiber as α, then $\alpha_k' \prec_C \alpha$ or α_k' and α are incomparable. Therefore, $\overrightarrow{g}_{ij}^k$ translated by $\nu \in \mathbb{N}^n$ is an edge of this graph, and we can move along it from α to a point α' in the same fiber such that $\alpha' \prec_C \alpha$ or α and α' are incomparable. This proves that, from every dominated point in the fiber, we can reach an improved or incomparable point (with respect to $\prec_C$) in the same fiber by moving along an edge of the graph. ∎

We call the graph in the b-fiber of $MIP_{A,C}$ built from elements in $\mathcal{G}_{\prec_C}$ the $\prec_C$-skeleton of that fiber.

The reader may note that, from each dominated solution α, one can easily build paths to its comparable Pareto-optimal solutions subtracting elements in $\mathcal{G}_{\prec_C}$. Indeed, let β a Pareto-optimal solution in the $A\alpha$-fiber such that β dominates α. Then, let α_i be a minimal element of $\mathcal{L}_C$ such that $\alpha = \alpha_i + \gamma$, with $\gamma \in \mathbb{N}^n$, and let β_i be the Pareto-optimal solution in the $A\alpha_i$-fiber that is comparable to α_i and such that $\beta_i + \gamma$ is comparable to β. Then $\alpha' = \beta_i + \gamma$ is a solution in the $A\alpha$-fiber with $\beta \prec_C \alpha' \prec_C \alpha$. Now, one repeats this process but starting with α' and β, until $\alpha' = \beta$. Moreover, the case where α and β are incomparable reduces to the previous one by finding a path from α to any intermediate point β' that compares with β. This analysis leads us to the following result.

COROLLARY 2.1. *In the $\prec_C$-skeleton of a fiber, there exists a directed path from every feasible point α to each Pareto-optimal point β in the same fiber. The vectors of objective function values of successive points in the path do not increase componentwise from α to β.*

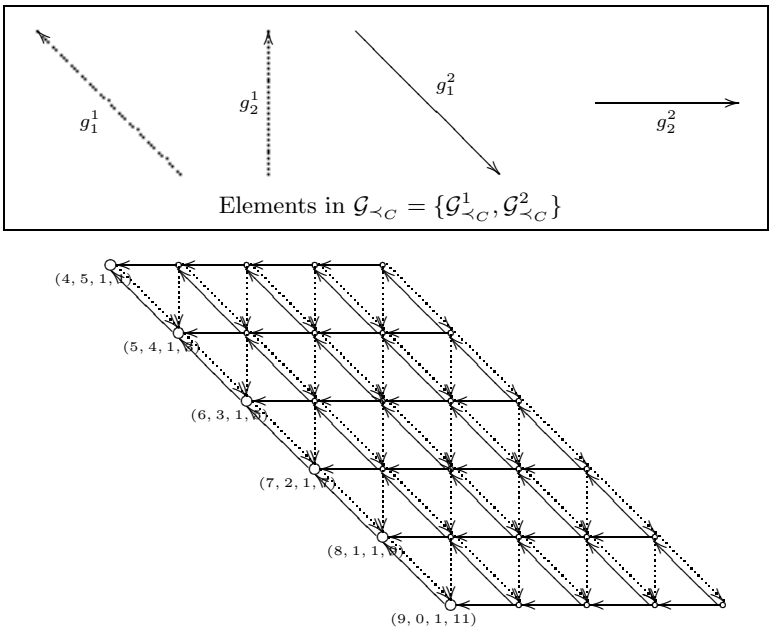


FIG. 2. *The $\prec_C$-skeleton of the $(17, 11)^t$-fiber of $MIP_{A,C}$ projected on the (x_1, x_2)-plane.*

COROLLARY 2.2. *The family $\mathcal{G}_{\prec_C}$ is the unique minimal test family for $MIP_{A,C}$. It depends only on the matrix A and the cost matrix C.*

Proof. By the definition of $\mathcal{G}_{\prec_C}$, conditions 1 and 2 of Definition 2.1 are satisfied. From Theorem 2.1, it follows that conditions 3 and 4 are also satisfied, so $\mathcal{G}_{\prec_C}$ is a test family for $MIP_{A,C}$. Minimality is due to the fact that removing any element (g_{ij}^k, h_{ij}^k) from $\mathcal{G}_{\prec_C}$ results in $\mathcal{G}_{\prec_C} \setminus \{(g_{ij}^k, h_{ij}^k)\}$. However, this new set is not a test family, since no oriented vector in $\mathcal{G}_{\prec_C} \setminus \{(g_{ij}^k, h_{ij}^k)\}$ can be translated through a nonnegative vector in $\mathbb{N}^n$ such that its tail meets α_j^i. It is clear by definition that $\mathcal{G}_{\prec_C}$ depends only on A and C. ∎

Example 2.2. Let $MIP_{A,C}$ be the family of multiobjective problems, with the following constraints and objective function matrices:

$$A = \begin{bmatrix} 2 & 2 & -1 & 0 \\ 0 & 2 & 0 & 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 10 & 1 & 0 & 0 \\ 1 & 10 & 0 & 0 \end{bmatrix}.$$

Let (x_1, x_2, s_1, s_2) be the vector of variables, where s_1 and s_2 are slack variables. In this example, using order $\prec_C^s$ (see Remark 2.2), $\mathcal{G}_{\prec_C} = \{\mathcal{G}^1_{\prec_C}, \mathcal{G}^2_{\prec_C}\}$, where $\mathcal{G}^1_{\prec_C} = \{\overrightarrow{g}^1_1 = ((0,1,2,-1),(0,1,2,0)),\ \overrightarrow{g}^1_2 = ((-1,1,0,-2),(0,1,0,0))\}$, and $\mathcal{G}^2_{\prec_C} = \{\overrightarrow{g}^2_1 = ((1,0,2,0),\ (1,0,2,0)),\ \overrightarrow{g}^2_2 = ((1,-1,0,2),(1,0,0,2))\}$.

Figure 2 shows, on the (x_1, x_2)-plane, the $\prec_C$-skeleton of the fiber corresponding to the right-hand side vector $(17, 11)^t$. In the box over the graph of the $\prec_C$-skeleton, we show the second components of the elements of $\mathcal{G}_{\prec_C}$. The reader may note that, in the graph, the arrows have opposite directions due to the fact that the directed paths (improving solutions) are built subtracting the elements in $\mathcal{G}_{\prec_C}$. We describe how to compute the sets $\mathcal{G}^1_{\prec_C}$ and $\mathcal{G}^2_{\prec_C}$ in section 3.

Given $\mathcal{G}_{\prec_C}$, there are several ways to build a path from each feasible point in a fixed fiber to any Pareto-optimal solution. However, there is a canonical way to do

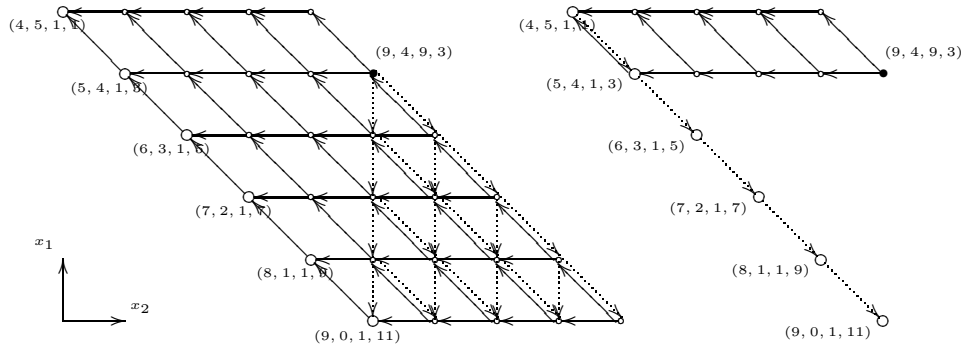


FIG. 3. *Different ways to compute paths from* $(9, 4, 9, 3)$ *to the Pareto-optimal solutions in its fiber.*

it: Fix σ a permutation of the set $\{1, \ldots, \mu\}$ and subtract from the initial point the elements of $\mathcal{G}_{\prec_C}^{\sigma(i)}$, for $i = 1, \ldots, \mu$. Add this element to an empty list. After each subtraction by elements in $\mathcal{G}_{\prec_C}^{\sigma(i)}$, $i = 1, \ldots, \mu$, remove from the list those elements dominated by the new element. We prove in Section 3 that this result does not depend on the permutation σ.

Example 2.2 (continuation). This example shows the above-mentioned different ways to compute paths from dominated solutions to any Pareto-optimal solution. The vector $(9, 4, 9, 3)$ is a feasible solution for $MIP_{A,C}$ in the $(17, 11)^t$-fiber. Figure 3 shows the sequence of Pareto-optimal points obtained from the feasible point $(9, 4, 9, 3)$ using the permutation $\sigma_1 = (1, 2)$ (on the left) and using $\sigma_2 = (2, 1)$ (on the right).

Remark 2.3. Let $\prec_C$ be the partial order induced by C. Then, a directed path from a dominated point α to each Pareto-optimal point β in a fiber, applying the above method, cannot pass through any lattice point in this fiber more than μ times (recall that μ is the number of maximal chains in $\mathcal{G}_{\prec_C}$). This implies that obtaining the Pareto-optimal solutions of a given $MIP_{A,C}$ using $\mathcal{G}_{\prec_C}$ cannot cycle.

3. Test families and partial Gröbner bases. In the previous section, we motivated the importance of having a test family for $MIP_{A,C}$, since this structure allows us to obtain the entire set of Pareto-optimal solutions of the above family of multiobjective integer programs (when the right-hand side varies). Our goal in this section is to provide the necessary tools to construct test families for any multiobjective integer problem. Our construction builds upon an extension of Gröbner bases on partial orders.

In order to introduce this structure, we define the reduction of a pair $(g, h) \in \mathbb{Z}^n \times \mathbb{Z}_+^n$ by a finite set of ordered pairs in $\mathbb{Z}^n \times \mathbb{Z}_+^n$. Given is a collection $\mathcal{G}_C \subseteq \mathbb{Z}^n \times \mathbb{Z}_+^n$, where $\mathcal{G}_C = \{(g_1, h_1), \ldots, (g_l, h_l) : h_{k+1} \prec_C h_k, k = 1, \ldots, l-1\}$.

The reduction of (g, h) by $\mathcal{G}_C$ consists of the process described in Algorithm 1. The above reduction process extends to the case of a finite collection of ordered sets of pairs in $\mathbb{Z}^n \times \mathbb{Z}_+^n$ by establishing the sequence in which the sets of pairs are considered. We denote by $pRem((g, h), \mathcal{G})_\sigma$ the reduction of the pair (g, h) by the family $\mathcal{G} = \{\mathcal{G}_i\}_{i=1}^t$ for a fixed sequence of indices σ.

From now on, we denote by $pRem((g, h), \mathcal{G})$ the set of remainders of (g, h) by the family $\mathcal{G} = \{\mathcal{G}_i\}_{i=1}^t$ for the natural sequence of indices $(1, \ldots, t)$, i.e., when σ is the identity.

ALGORITHM 1: PARTIAL REDUCTION ALGORITHM.

input : $R = \{(g,h)\}$, $S = \{(g,h)\}$,
$\mathcal{G}_C = \{(g_1,h_1),\ldots,(g_l,h_l) : h_{k+1} \prec_C h_k, k = 1,\ldots,l-1\}$.
Set $i := 1$, $S_o = \{\}$.
repeat
  for $(\tilde{g},\tilde{h}) \in S \setminus S_o$ **do**
    while $\tilde{h} - h_i \geq 0$ **do**
      if $\tilde{h} - g_i$ and $\tilde{h} - \tilde{g}$ are comparable by $\prec_C$ **then**
        $R_o = \{(\tilde{g} - g_i, \max_{\prec_C}\{\tilde{h} - g_i, \tilde{h} - \tilde{g}\})\}$
      else
        $R_o = \{(\tilde{g} - g_i, \tilde{h} - g_i), (\tilde{g} - g_i, \tilde{h} - \tilde{g})\}$
      end
      For each $r \in R_o$ and $s \in R$:
      if $r \prec_C s$ **then**
        $R = R \backslash \{s\}$;
      end
      $S = R_o$.
      $R = R \cup R_o$.
      $S_o = S_o \cup \{(\tilde{g},\tilde{h})\}$.
    end
  end
  $i = i + 1$.
until $i \leq t$;
output: R, the partial reduction set of (g,h) by $\mathcal{G}_C$.

The reduction of a pair that represents a feasible solution by a test family gives the entire set of Pareto-optimal solutions. In order to obtain that test family, we introduce the notion of p-Gröbner basis. This name has been motivated by the fact that when the ordering in $\mathbb{N}^n$ is induced by a single cost vector, a Gröbner basis is a test set for the family of integer programs $IP_{A,c}$ (see [10] or [41] for extended details). In the single objective case, the Buchberger algorithm computes a Gröbner basis. However, in the multiobjective case, the cost matrix induces a partial order, so division or the Buchberger algorithm are not applicable. Using the above reduction algorithm (Algorithm 1), we present an "á la Buchberger" algorithm to compute the so called p-Gröbner basis to solve MOILP problems.

DEFINITION 3.1 (partial Gröbner basis). *A family $\mathcal{G} = \{\mathcal{G}_1,\ldots,\mathcal{G}_t\} \subseteq I_A$ is a partial Gröbner basis (p-Gröbner basis) for the family of problems $MIP_{A,C}$ if $\mathcal{G}_1,\ldots,\mathcal{G}_t$ are the maximal chains for the partially ordered set $\bigcup_{i=1}^{t}\mathcal{G}_i$ and for any $(g,h) \in \mathbb{Z}^n \times \mathbb{Z}^n_+$, with $h - g \geq 0$*

$$g \in \mathrm{Ker}(A) \Longleftrightarrow pRem((g,h),\mathcal{G})_\sigma = \{0\}$$

for any sequence σ.

A p-Gröbner basis is said to be reduced if every element in each maximal chain cannot be obtained by reducing any other element of the same chain.

Given a p-Gröbner basis, computing a reduced p-Gröbner basis is done by deleting the elements that can be reduced by other elements in the basis. After the removing process, the family is a p-Gröbner basis having only nonredundant elements. It is easy to see that the reduced p-Gröbner basis for $MIP_{A,C}$ is unique and minimal, in the sense that no element can be removed from it, maintaining the p-Gröbner basis structure.

This definition clearly extends to p-Gröbner bases for the ideal I_A induced by A, once we fix the partial order $\prec_C$ induced by C.

In the following, we present algorithms to solve multiobjective problems analogous to the methods that solve the single objective case using usual Gröbner basis. These methods are based on computing the reduction of a feasible solution by the basis. The key for that result is the fact that the reduction of any pair of feasible solutions is the same, therefore, the algorithm is valid for any initial feasible solution. After the following theorem, Lemma 3.1 ensures the same statement for the multiobjective case and p-Gröbner bases.

THEOREM 3.1. *Let $\mathcal{G}$ be the reduced p-Gröbner basis for $MIP_{A,C}$ and α a feasible solution for $MIP_{A,C}(A\alpha)$. Then, $pRem((\alpha,\alpha),\mathcal{G})_\sigma = pRem((\alpha,\alpha),\mathcal{G})_{\sigma'}$ for any sequences σ and σ'.*

Proof. We first observe that the elements in $pRem((\alpha,\alpha),\mathcal{G})_\sigma$ are of the form (β,β). Indeed, since the first step of Algorithm 1 reduces the element (α,α), then $\tilde{h}-\tilde{g}=\alpha-\alpha=0$. Therefore, $\tilde{h}-\tilde{g}$ is always dominated by $\tilde{h}-g_i$ because $0 \prec_C \tilde{h}-g_i$, so that the remainders are of the form $(\alpha-g_i,\alpha-g_i)$.

On other hand, let (β,β) be an element in $pRem((\alpha,\alpha),\mathcal{G})_\sigma$, then $\alpha-\beta \in \mathrm{Ker}(A)$ and by Definition 3.1, $pRem((\alpha-\beta,\alpha),\mathcal{G})_{\sigma'} = pRem((\alpha-\beta,\beta),\mathcal{G})_{\sigma'} = \{0\}$ for any σ'. ∎

The above result ensures that without loss of generality reductions of elements of the form (α,α) by p-Gröbner bases are independent of the permutation of indices used. Therefore, we do not make reference to σ in the notation, referring always to the natural sequence $\sigma=(1,\ldots,t)$.

LEMMA 3.1. *Let $\mathcal{G}$ be the reduced p-Gröbner basis for $MIP_{A,C}$ and α_1,α_2 two different feasible solutions in the same fiber of $MIP_{A,C}$. Then, $pRem((\alpha_1,\alpha_1),\mathcal{G}) = pRem((\alpha_2,\alpha_2),\mathcal{G})$.*

Proof. Let $(\beta,\beta) \in pRem((\alpha_1,\alpha_1),\mathcal{G})$, then since $A\alpha_1 = A\alpha_2$, β is in the same fiber that α_2. Next, since β cannot be reduced, then $(\beta,\beta) \in pRem((\alpha_2,\alpha_2),\mathcal{G})$. ∎

The following theorem states the relationship between the three structures introduced before: test families, reduced p-Gröbner bases, and the family $\mathcal{G}_{\prec_C}$.

THEOREM 3.2. *The reduced p-Gröbner basis for $MIP_{A,C}$ is the unique minimal test family for $MIP_{A,C}$. Moreover, $\mathcal{G}_{\prec_C}$, introduced in* (3)*, is the reduced p-Gröbner basis for $MIP_{A,C}$.*

Proof. Let $\mathcal{G}=\{\mathcal{G}_1,\ldots,\mathcal{G}_t\}$ be the reduced p-Gröbner basis for $MIP_{A,C}$. We have to prove that $\mathcal{G}$ satisfies the four conditions in Definition 2.1. By the definition of p-Gröbner basis, it is clear that each $\mathcal{G}_i$ is totally ordered by its second component with respect to $\prec_C$ (condition 1). Condition 2 follows because, for each i and for each $(g,h) \in \mathcal{G}_i \subseteq \mathbb{Z}^n \times \mathbb{Z}^n_+$, clearly $pRem((g,h),\mathcal{G}) = \{0\}$, so $g \in \mathrm{Ker}(A)$ and then $A(h-g)=Ah$.

Now, let $x \in \mathbb{N}^n$ be a dominated solution for $MIP_{A,C}(b)$. Then, there is a Pareto-optimal solution β such that $\beta \prec_C x$. By Lemma 3.1, $pRem((x,x),\mathcal{G}) = pRem((\beta,\beta),\mathcal{G})$ and by construction of the set of partial remainders, $\beta \in pRem((\beta,\beta),\mathcal{G})$, thus $x \notin pRem((x,x),\mathcal{G})$. This implies that there exists $(g,h) \in \mathcal{G}_i$, for some $i=1,\ldots,t$ such that $x-g \prec_C x$. This proves condition 3 of Definition 2.1.

On the other hand, if x is a Pareto-optimal solution for $MIP_{A,C}(b)$, $x \in pRem((x,x),\mathcal{G})$, then there exists no (g,h) in any $\mathcal{G}_i$ such that $x-g \prec_C x$. Therefore, for every i and for each $(g,h) \in \mathcal{G}_i$, either $x-g$ is infeasible or incomparable to x, which proves condition 4 of Definition 2.1.

Minimality is due to the fact that removing an element from the reduced p-Gröbner basis, that is, the minimal partial Gröbner basis that can be built for $MIP_{A,C}$,

we cannot guarantee to have a test family because there may exist a pair $(g,h) \in \mathbb{Z}^n \times \mathbb{Z}^n_+$, with $g \in \mathrm{Ker}(A)$ that cannot be reduced to the zero set.

Finally, the second statement of the theorem follows from Corollary 2.2. ∎

Next, we describe an extended algorithm to compute a p-Gröbner basis for I_A, with respect to the partial order induced by C. First, for any $(g,h),(g',h')$ in $\mathbb{Z}^n \times \mathbb{Z}^n_+$, we denote by $S^1((g,h),(g',h'))$ and $S^2((g,h),(g',h'))$ the pairs

$$S^1((g,h),(g',h')) = \begin{cases} (g-g'-2(h-h'),\gamma+g-2h) & \text{if } \gamma+g-2h \prec_C \gamma+g'-2h', \\ (g'-g-2(h'-h),\gamma+g'-2h') & \text{if } \gamma+g'-2h' \prec_C \gamma+g-2h, \\ (g-g'-2(h-h'),\gamma+g-2h) & \text{if } \gamma+g'-2h' \text{ and } \gamma+g-2h, \\ & \text{are incomparable,} \end{cases}$$

and

$$S^2((g,h),(g',h')) = \begin{cases} (g-g'-2(h-h'),\gamma+g-2h) & \text{if } \gamma+g-2h \prec_C \gamma+g'-2h', \\ (g'-g-2(h'-h),\gamma+g'-2h') & \text{if } \gamma+g'-2h' \prec_C \gamma+g-2h, \\ (g'-g-2(h'-h),\gamma+g'-2h') & \text{if } \gamma+g'-2h' \text{ and } \gamma+g-2h, \\ & \text{are incomparable,} \end{cases}$$

where $\gamma \in \mathbb{N}^n$ and $\gamma_i = \max\{h_i, h_i'\}$, $i = 1,\ldots,n$.

The pairs $S^1((g,h),(g',h'))$ and $S^2((g,h),(g',h'))$ are called $1-Svector$ and $2-Svector$ of (g,h) and (g',h'), respectively. The reader may note that $S^1((g,h),(g',h'))$ and $S^2((g,h),(g',h'))$ coincide, provided that the resulting pairs are comparable under $\prec_C$, whereas they correspond with the two possible choices of the new pair in the case when the vectors $\gamma+g'-2h'$ and $\gamma+g-2h$ are incomparable.

The name is due to the analogy with the algebraic-geometrical notion of S-polynomial for a pair of polynomials with a given term order. Since we consider a partial order, it may happen that in the standard construction of an Svector [41], we cannot decide which is the leading term. Therefore, in our definitions of Svectors, we must consider all possible combinations of leading terms, with respect to the partial order $\prec_C$.

The original Buchberger criterion was stated in a polynomial language. Therefore, we adapt our notation to follow the line of that proof. Let $leadmon_C(f)$ denote the set of leading monomials with respect to the order induced by C for any multivariate polynomial $f \in \mathbb{R}[x_1,\ldots,x_n]$. We identify the set J_A introduced in (2), with $\Im_A = \langle x^u - x^v : u - v \in \mathrm{Ker}(A)\rangle$, and therefore, the set $setlm(u,v)$ is identified with the elements in $leadmon_C(x^u - x^v)$. Moreover, each pair $(g,h) \in \mathbb{Z}^n \times \mathbb{Z}^n_+$, with $g \in \mathrm{Ker}(A)$ and $h - g \geq 0$ is identified with the binomial $x^h - x^{h-g}$. Then, we associate with $\mathcal{G} = \{\mathcal{G}_1,\ldots,\mathcal{G}_t\}$ the polynomial set $\mathcal{G}^* = \{\mathcal{G}_1^*,\ldots,\mathcal{G}_t^*\}$ identifying one–to-one each pair in $\mathcal{G}$ with the corresponding binomial in $\mathcal{G}^*$. In this way, we adapt accordingly the definition of $pRem((f,p),\mathcal{G}^*)$, the set of partial remainders of $f \in \mathbb{R}[x_1,\ldots,x_n]$ with leading monomial p and with respect to $\mathcal{G}^*$.

Moreover, we define the 1-Spolynomial and 2-Spolynomial as the binomial transcriptions of the 1-Svector and 2-Svector. For any two binomials $x^{\alpha_1} - x^{\beta_1}$ and $x^{\alpha_2} - x^{\beta_2}$, the k-Spolynomial with respect to the leading monomials x^{α_1}, x^{α_2} is

$$\mathbb{S}^k((x^{\alpha_1} - x^{\beta_1}, x^{\alpha_1}),(x^{\alpha_2} - x^{\beta_2}, x^{\alpha_2})) = x^{\gamma-\alpha_2+\beta_2} - x^{\gamma-\alpha_1+\beta_1}, \qquad k = 1,2,$$

where $\gamma \in \mathbb{N}^n$ and $\gamma_i = \max\{(\alpha_1)_i,(\alpha_2)_i\}$, $i = 1,\ldots,n$. The difference between the 1-Spolynomial and the 2-Spolynomial is the choice of the leading term: They coincide when the monomials are comparable and differ when the monomials are incomparable,

and in this case, each k-Spolynomial corresponds with the two possible choices of the leading term.

The following lemma is used in the proof of our extended criterion, and it is an adaptation of the analogous result for total orders and usual S-polynomials.

LEMMA 3.2. *Let $f_1, \ldots, f_s \in \mathbb{R}[x_1, \ldots, x_n]$ be such that there exists $p \in \bigcap_{i=1}^{s} leadmon_C(f_i)$. Let $f = \sum_{i=1}^{s} c_i\, f_i$, with $c_i \in \mathbb{R}$. If there exists $q \in leadmon_C(f)$ such that $q \prec_C p$, then f is a linear combination with coefficients in $\mathbb{R}$ of the k-Spolynomial, $k = 1, 2$, of f_i and f_j, $1 \leq i < j \leq s$.*

Proof. By hypothesis, $f_i = a_i\, p +$ *other smaller or incomparable terms*, with $a_i \in \mathbb{R}$ for all i. Then, f can be rewritten as $f = \sum_{i=1}^{s} c_i\, f_i = \sum_{i=1}^{s} c_i\, a_i\, p +$ *other smaller or incomparable terms*. Since $q \prec_C p$, then $\sum_{i=1}^{s} c_i\, a_i = 0$.

By definition, for $k = 1, 2$, $\mathbb{S}^k((f_i, p), (f_j, p)) = \frac{1}{a_i} f_i - \frac{1}{a_j} f_j$, thus,

$$
\begin{aligned}
f &= c_1\, f_1 + \cdots + c_s\, f_s = c_1\, a_1 \left(\frac{1}{a_1} f_1\right) + \cdots + c_s\, a_s \left(\frac{1}{a_s} f_s\right) \\
&= c_1\, a_1 \left(\frac{1}{a_1} f_1 - \frac{1}{a_2} f_2\right) + (c_1\, a_1 + c_2\, a_2) \left(\frac{1}{a_2} f_2 - \frac{1}{a_3} f_3\right) + \cdots \\
&+ (c_1\, a_1 + \cdots + c_{s-1}\, a_{s-1}) \left(\frac{1}{a_{s-1}} f_{s-1} - \frac{1}{a_s} f_s\right) + (c_1\, a_1 + \cdots + c_s\, a_s) \frac{1}{a_s} f_s \\
&= d_1^k\, \mathbb{S}^k((f_1, p), (f_2, p)) + \cdots + d_{s-1}^k\, \mathbb{S}^k((f_{s-1}, p), (f_s, p)) + \left(\frac{1}{a_s} \sum_{i=1}^{s} c_i\, a_i\right) f_s \\
&= \sum_{i=1}^{s-1} d_i^k\, \mathbb{S}^k((f_i, p), (f_{i-1}, p)),
\end{aligned}
$$

where $d_i^k = \sum_{j=1}^{i} c_j\, a_j$ for $i = 1, \ldots, s$ and $k = 1, 2$. This proves the lemma. ∎

The algorithm to compute standard Gröbner bases is based on the Buchberger criterion. Its analogous for a partial order states that it suffices to check that the partial remainders are zero for Svectors and for any fixed sequence of indices.

THEOREM 3.3 (extended Buchberger's criterion). *Let $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_t\}$, with $\mathcal{G}_i \subseteq I_A$ for all $i = 1, \ldots, t$, be the maximal chains of the partially ordered set $\{g_i : g_i \in \mathcal{G}_i$ for some $i = 1, \ldots, t\}$ and such that $\mathcal{G}^*$, the polynomial transcription of $\mathcal{G}$, is a system of generators of $\Im_A$. Then the following statements are equivalent:*

(1) *$\mathcal{G}$ is a p-Gröbner basis for the family $MIP_{A,C}$.*

(2) *For each $i, j = 1, \ldots, t$ and $(g, h) \in \mathcal{G}_i$, $(g', h') \in \mathcal{G}_j$, $pRem(S^k((g, h), (g', h')), \mathcal{G}) = \{0\}$ for $k = 1, 2$.*

Proof. Let $\mathcal{G}$ be a p-Gröbner basis for I_A and $(g, h) \in \mathcal{G}_i$, $(g', h') \in \mathcal{G}_j$ for any $i, j \in \{1, \ldots, t\}$. Then, $S^k((g, h), (g', h'))$, for $k = 1, 2$, is in I_A so by the definition of p-Gröbner basis, $pRem(S^k((g, h), (g', h')), \mathcal{G})_\sigma = \{0\}$, for any sequence σ, in particular, for $\sigma = (1, \ldots, t)$.

Conversely, assume that for each $(g, h) \in \mathcal{G}_i$ and $(g', h') \in \mathcal{G}_j$ for any $i, j \in \{1, \ldots, t\}$, $pRem(S^k((g, h), (g', h')), \mathcal{G}) = \{0\}$ for $k = 1, 2$. Let $(\tilde{g}, \tilde{h}) \in \mathbb{Z}^n \times \mathbb{Z}_+^n$, with $\tilde{g} \in \mathrm{Ker}(A)$ and $\tilde{h} - \tilde{g} \geq 0$. We define $f = x^{\tilde{h}} - x^{\tilde{h} - \tilde{g}} \in \mathbb{Z}[x_1, \ldots, x_n]$, and we denote by $\mathcal{G}^* = \{g_1^*, \ldots, g_d^*\}$ the polynomial set associated with $\mathcal{G}$.

Then, by hypothesis f, can be written as a linear combinations of the elements in $\mathcal{G}^*$ (this representation is not unique):

$$f = \sum_{i=1}^{d} p_i\, g_i^*$$

for some $p_i \in \mathbb{R}[x_1, \ldots, x_n]$ for $i = 1, \ldots, d$.

Let $X = \{X_1, \ldots, X_N\}$ be the set of maximal elements of the set $\{P_i\, R_i : P_i \in leadmon_C(p_i),\ R_i \in leadmon_C(g_i^*)\}$ with respect to $\prec_C$.

If $X \supseteq leadmon_C(f)$, the polynomial f can be partially reduced by the elements in $\mathcal{G}^*$. This proves the result.

Otherwise, there must exist $l \in leadmon_C(f) \backslash X$. We will prove by contradiction that this case is not possible. Indeed, if $l \in leadmon_C(f)$, it must come from some simplification (reduction) of the linear combination defining f. Then, the construction ensures that there must exist at least one element $X_i \in X$ such that $l \prec_C X_i$.

Set $J(X_i) = \{j : P_j\, R_j = X_i, \text{ with } P_j \in leadmon_C(p_j), R_j \in leadmon_C(g_j^*)\}$. For any $j \in J(X_i)$, we can write $p_j = P_j + \ \textit{other terms}$ and define $q = \sum_{j \in J(X_i)} P_j\, g_j^*$. Then, $X_i \in leadmon_C(P_j\, g_j^*)$ for all $j \in J(X_i)$. However, by hypothesis, there exists $Q \in leadmon_C(q)$, with $Q \prec_C X_i$.

Hence, by Lemma 3.2, there exist $d_{s,r}^k \in \mathbb{R}$, $k = 1, 2$ such that

$$q = \sum_{r,s \in J(X_i), r \neq s, g_s^*, g_r^* \in \mathcal{G}^*} d_{s,r}^k\, \mathbb{S}^k((P_s\, g_s^*, L_s), (P_r\, g_r^*, L_s)), \quad k = 1, 2$$

for some $L_j \in leadmon_C(P_j\, g_j^*)$ for all $g_j^* \in \mathcal{G}^*$.

Now, for any r, $s \in J(X_i)$, we have that $X_i = lcm(L_r, L_s)$ for some $L_r \in leadmon_C(P_r\, g_r^*)$ and $L_s \in leadmon_C(P_s\, g_s^*)$, and therefore, we can write

$$\begin{aligned}\mathbb{S}^k((P_r\, g_r^*, L_r), (P_s\, g_s^*, L_s)) &= \frac{X_i}{L_r}\, P_r\, g_r^* - \frac{X_i}{L_s}\, P_s\, g_s^* \\ &= \frac{X_i}{l_r}\, g_r^* - \frac{X_i}{l_s}\, g_s^* = \frac{X_i}{P_{r,s}}\, \mathbb{S}^k((g_r^*, l_r), (g_s^*, l_s)),\end{aligned}$$

where $l_r = \frac{L_r}{P_r}$, $l_s = \frac{L_r}{P_s}$, $P_{r,s} = lcm(l_r, l_s)$, and $k = 1, 2$.

By hypothesis, $pRem(\mathbb{S}^k((g_r^*, l_r), (g_s^*, l_s)), \mathcal{G}^*) = \{0\}$. Thus, from the last equation we deduce that

$$pRem(\mathbb{S}^k((P_r\, g_r^*, L_r), (P_s\, g_s^*, L_s)), \mathcal{G}) = \{0\}.$$

This gives a representation:

$$\mathbb{S}^k((P_r\, g_r^*, L_r), (P_s\, g_s^*, L_s)) = \sum_{g_\nu^* \in \mathcal{G}^*} p_{r,s}^{k,\nu}\, g_\nu^*,$$

with $p_{r,s}^{k,\nu} \in \mathbb{R}[x_1, \ldots, x_n]$ and $k = 1, 2$.

Then, $\{P_{r,s}^{k,\nu}\, R^\nu\colon\ g_\nu^* \in \mathcal{G}^*, P_{r,s}^{k,\nu} \in leadmon_C(p_{r,s}^{k,\nu}), R^\nu \in leadmon_C(g_\nu^*)$ and do not exist $P_{r,s}^{k,\tilde{\nu}}$ and $R^{\tilde{\nu}}$ satisfying $P_{r,s}^{k,\tilde{\nu}} \in leadmon_C(p_{r,s}^{k,\tilde{\nu}})$, $R^{\tilde{\nu}} \in leadmon_C(g_{\tilde{\nu}}^*)$ such that $P_{r,s}^{k,\nu}\, R^\nu \prec_C P_{r,s}^{k,\tilde{\nu}}\, R^{\tilde{\nu}}\} = leadmon_C(\mathbb{S}^k(P_r\, g_r^*, P_s\, g_s^*))$.

To simplify the notation, denote $S_{r,s}^k = leadmon_C(\mathbb{S}^k(P_r\, g_r^*, P_s\, g_s^*))$.

By construction of S-polynomials, we have that there exists $p \in S^k_{r,s}$ such that $p \prec_C X_i$, so, substituting these expressions into q above, we have

$$f = \sum_{j \notin J(X_i)} p_j\, g_j^* + \sum_{j \in J(X_i)} p_j\, g_j^* = \sum_{j \notin J(X_i)} p_j\, g_j^* + q$$
$$= \sum_{j \notin J(X_i)} p_j\, g_j^* + \sum_{r,s} d^k_{r,s}\, \mathbb{S}^k((P_s\, g_s^*, L_s), (P_r\, g_r^*, L_r)) = \sum_{j \notin J(X_i)} p_j\, g_j^* + \sum_{r,s} \sum_{\nu} p^{k,\nu}_{r,s}\, g_\nu^*.$$

Thus, we have expressed f as

$$f = \sum_{i=1}^{d} p_i'\, g_i^*,$$

with one leading term p smaller than X_i. However, this is a contradiction, and the theorem is proved. ∎

This criterion (the one in Theorem 3.3) allows us to describe a geometric algorithm which constructs a p-Gröbner basis $\mathcal{G}_C$ for $MIP_{A,C}$, and therefore, a test family for that family of multiobjective problems.

The first approach to compute a p-Gröbner basis for a family of multiobjective programs is based on the Conti and Traverso method for the single objective case [10]. For this algorithm, the key is transforming the given multiobjective program into another one where computations are easier and so that an initial set of generators for I_A is known.

Notice that finding an initial set of generators for I_A can be done by a straightforward modification of the Big-M method [3].

Given the program $MIP_{A,C}(b)$, we consider the associated extended multiobjective program $EMIP_{A,C}(b)$ as the problem $MIP_{\widetilde{A},\widetilde{C}}(b)$, where

$$\widetilde{A} = \left(\begin{array}{c|c|c} & -1 & \\ Id_m & \vdots & A \\ & -1 & \end{array} \right) \in \mathbb{Z}^{m \times (m+1+n)},$$

$\widetilde{C} = (M \cdot \mathbf{1} | C) \in \mathbb{Z}^{(m+1+n)\times k}$, Id_m stands for the $m \times m$ identity matrix, M is a large constant, and $\mathbf{1}$ is the $(m+1) \times k$ matrix whose components are all 1. This problem adds $m+1$ new variables, whose weights in the multiobjective function are big, and so solving this extended minimization program allows us to solve directly the initial program $MIP_{A,C}$. Indeed, any feasible solution to the original problem is a feasible solution to the extended problem with the first m components equal to zero, so any feasible solution of the form $(0, \overset{m+1}{\ldots}, 0, \alpha_1, \ldots, \alpha_n)$ is nondominated, upon the order $\prec_{\widetilde{C}}$, by any solution without zeros in the first m components. Then, computing a p-Gröbner basis for the extended program using the partial Buchberger Algorithm (Algorithm 2) allows detecting infeasibility of the original problem. Furthermore, a trivial feasible solution $\widetilde{\mathbf{x}}_0 = (b_1, \ldots, b_m, 0, \overset{n+1}{\ldots}, 0)$ is known, and the initial set of generators for I_A is given by $\{\{M_i - P_i, M_i\} : i = 0 \ldots, n\}$, where $M_i = (a_{1i} - \min\{0, \min_j\{a_{ji}\}\}, \ldots, a_{mi} - \min\{0, \min_j\{a_{ji}\}\}, -\min\{0, \min_j\{a_{ji}\}\}, 0, \overset{n}{\ldots}, 0)$, $P_i = (0, \overset{m+1}{\ldots}, 0 | e_i)$, for all $i = 1, \ldots, n$, $M_0 = (1, \overset{m+1}{\ldots}, 1, 0, \overset{n}{\ldots}, 0)$, and $P_0 = \mathbf{0}$, $M_i, P_i, M_0, P_0 \in \mathbb{Z}^{n+m+1}_+$ (see [2] for further details). Then, we can state the following result.

THEOREM 3.4. *Let $\mathcal{G} = \{\mathcal{G}_i\}_{i=1}^t$ be a p-Gröbner basis for $EMIP_{A,C}$ and $\mathbf{b} = (b_1, \ldots, b_m)$. The entire set of Pareto-optimal solutions for $MIP_{A,C}(\mathbf{b})$ consists*

ALGORITHM 2: PARTIAL BUCHBERGER ALGORITHM I.

input : $F_1 = \{M_0, M_1, \ldots, M_n\}$ and $F_2 = \{P_0, P_1, \ldots, P_n\}$,
$M_i = (a_{1i} - \min\{0, \min_j\{a_{ji}\}\}, \ldots, a_{mi} - \min\{0, \min_j\{a_{ji}\}\}, -\min\{0, \min_j\{a_{ji}\}\}, 0, \overset{n}{\ldots}, 0)$ $(i > 0)$,
$P_i = (0, \overset{m+1}{\ldots}, 0|e_i) \in \mathbb{N}^{m+n+1}$ $(i > 0)$,
$M_0 = (1, \overset{m+1}{\ldots}, 1, 0, \overset{n}{\ldots}, 0)$,
$P_0 = (0, \overset{n+m+1}{\ldots}, 0)$.

repeat
  Compute, $\mathcal{G}_1, \ldots, \mathcal{G}_t$, the maximal chains for $\mathcal{G} = \phi(\mathcal{F}(F_1, F_2))$.
  for $i, j \in \{1, \ldots, t\}$, $i \neq j$, *and each pair* $(g, h) \in \mathcal{G}_i$, $(g', h') \in \mathcal{G}_j$ **do**
    Compute $R^k = pRem(S^k((g, h), (g', h')), \mathcal{G})$, $k = 1, 2$.
    if $R^k = \{0\}$ **then**
      Continue with other pair.
    else
      Add $\phi(\mathcal{F}(r))$ to $\mathcal{G}$ for each $r \in R^k$.
    end
  end
until $R^k = \{0\}$ *for every pairs*;
output: $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_Q\}$ p-Gröbner basis for I_A with respect to $\prec_C$.

of the vectors $\alpha = (\alpha_1, \ldots, \alpha_n)$ *such that* $(0, \overset{m+1}{\ldots}, 0, \alpha_1, \ldots, \alpha_n) \in pRem(((\mathbf{b}, 0, \overset{n+1}{\ldots}, 0), (\mathbf{b}, 0, \overset{n+1}{\ldots}, 0)), \mathcal{G})$. *Moreover, if there is no* α' *in the set* $pRem(((\mathbf{b}, 0, \overset{n+1}{\ldots}, 0), (\mathbf{b}, 0, \overset{n+1}{\ldots}, 0)), \mathcal{G})$ *whose* $m+1$ *first components are zero* $MIP_{A,C}(b)$ *is infeasible.*

Proof. Let α be a vector obtained by successive reductions over $\mathcal{G}$. It is clear that α is feasible because $((\mathbf{0}, \alpha), (\mathbf{0}, \alpha))$ is in the set of remainders of $((\mathbf{b}, \mathbf{0}), (\mathbf{b}, \mathbf{0}))$ by $\mathcal{G}$ and then in the same fiber. Besides, α is a Pareto-optimal solution because $\mathcal{G}$ is a test family for the problem (Theorem 3.2).

Now, if β^* is a Pareto-optimal solution, by Lemma 3.1 $pRem(((\mathbf{0}, \beta^*), (\mathbf{0}, \beta^*)), \mathcal{G}) = pRem(((\mathbf{0}, \mathbf{b}), (\mathbf{0}, \mathbf{b})), \mathcal{G})$, but since β^* is a Pareto-optimal solution, it cannot be reduced so $((\mathbf{0}, \beta^*), (\mathbf{0}, \beta^*)) \in pRem(((\mathbf{0}, \beta^*), (\mathbf{0}, \beta^*)), \mathcal{G})$ and then $((\mathbf{0}, \beta^*), (\mathbf{0}, \beta^*))$ also belongs to the list of partial remainders of $((\mathbf{b}, \mathbf{0}), (\mathbf{b}, \mathbf{0}))$ by $\mathcal{G}$. ∎

Hoşten and Sturmfels [22] improved the method by Conti and Traverso to solve single-objective programs using standard Gröbner bases. Their improvement is due to the fact that it is not necessary to increase the number of variables in the problem, as Conti and Traverso's algorithm does. Hoşten and Sturmfels's algorithm allows decreasing the number of steps in the computation of the Gröbner basis, but, on the other hand, it needs an algorithm to compute an initial feasible solution, which was trivial in the Conti and Traverso algorithm. We have modified this alternative algorithm to compute the entire set of Pareto-optimal solutions. The first step in the algorithm is computing an initial basis for the polynomial toric ideal $\Im_A = \langle x^u - x^v : u - v \in \mathrm{Ker}(A)\rangle$ that we can identify with J_A. This step does not depend on the order induced by the objective function, so it can be used to solve multiobjective problems. Details can be seen in [22]. Algorithm 3 implements the computation of the set of generators of $\Im_A$. This procedure uses the notion of Lenstra–Lenstra–Lovàsz (LLL)-reduced basis (see [27] for further details). In addition, we use a ω-graded reverse lexicographic term order $\prec_\omega^{gr_i}$ induced by $x_{i+1} > \cdots > x_{i-1} > x_i$ (with $x_{n+1} := x_1$) that is defined as follows:

$$\alpha \prec_\omega^{gr_i} \beta :\Longleftrightarrow \sum_{j=1}^{n} \omega_j \alpha_j < \sum_{j=1}^{n} \omega_j \beta_j \text{ or } \sum_{j=1}^{n} \omega_j \alpha_j = \sum_{j=1}^{n} \omega_j \beta_j \text{ and } \alpha \prec_{lex} \beta,$$

ALGORITHM 3: `setofgenerators(A)`.

input : $A \in \mathbb{Z}^{m\times n}$

1. Find a lattice basis $\mathcal{B}$ for $\mathrm{Ker}(A)$ (using Hermite normal form).
2. Replace $\mathcal{B}$ by the LLL-reduced lattice basis $\mathcal{B}_{red}$.
   Let $J_0 := \langle x^{u_+} - x^{u_-} : u \in \mathcal{B}_{red}\rangle$.
   for $i = 1, \ldots, n$ **do**
   Compute $J_i = (J_{i-1} : x_i^{\infty})$ as
   (a) Compute $\mathcal{G}_{i-1}$ the reduced Gröbner basis for J_{i-1} with respect to $\prec_{\omega}^{gr_i}$.
   (b) Divide each element $f \in \mathcal{G}_{i-1}$ by the highest power of x_i that divides f.

output: $\Im_A := J_n = \{x^{u_1} - x^{v_1}, \ldots, x^{u_s} - x^{v_s}\}$ system of generators for I_A.

ALGORITHM 4: `pgrobner(F1, F2)`.

input : $F_1 = \{M_1, \ldots, M_s\}$ and $F_2 = \{P_1, \ldots, P_s\}$.
repeat
  Compute $\mathcal{G}_1, \ldots, \mathcal{G}_t$ the maximal chains for $\mathcal{G} = \phi(\mathcal{F}(F_1, F_2))$.
  for $i, j \in \{1, \ldots, t\}$, $i \neq j$, *and each pair* $(g, h) \in \mathcal{G}_i$, $(g', h') \in \mathcal{G}_j$ **do**
    Compute $R^k = pRem(S^k((g, h), (g^{'}, h^{'})), \mathcal{G})$, $k = 1, 2$.
    if $R^k = \{0\}$ **then**
      Continue with other pair.
    else
      Add $\phi(\mathcal{F}(r))$ to $\mathcal{G}$ for each $r \in R^k$.
    end
  end
until $R^k = \{0\}$ *for every pairs*;
output: $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_Q\}$ p-Gröbner basis for $MIP_{A,C}$.

where $\omega \in \mathbb{R}_+^n$ is chosen such that $x_{i+1} > \cdots > x_{i-1} > x_i$. Finally, for any $a \in \mathbb{R}$, we denote by $a_+ = \max\{a, 0\}$ and $a_- = -\min\{a, 0\}$.

$\Im_A$ consists of binomials $x^{u_i} - x^{v_i}$ with $u_i - v_i \in \mathrm{Ker}(A)$ for $i = 1, \ldots, s$. Coming back to our notation, each binomial $x^u - x^v$ in $\Im_A$ is identified with $\{u, v\} \in J_A$, so computing a set of generators for $\Im_A$ gives us, in some sense, a finite number of generators for the set that represents the constraints matrix. We compute in the next step a partial Gröbner basis from initial sets $F_1 = \{u_1, \ldots, u_s\}$ and $F_2 = \{v_1, \ldots, v_s\}$ using our extended Buchberger algorithm (Algorithm 4).

Once we have obtained the partial Gröbner basis using the above algorithm, we can compute the entire set of Pareto-optimal solutions for $MIP_{A,C}(b)$ by Algorithm 5.

There are some interesting cases where our methodology is highly simplified due to the structure of the set of constraints. One of these cases is when the dimension of the set of constraints is $n-1$. The next remark explains how the algorithm simplifies in this case.

Remark 3.1. Let A be an $m \times n$ integer matrix with rank $n-1$. Then, since $\dim(\mathrm{Ker}(A)) = 1$, the system of generators for I_A (Step 2) has just one element, (g, h) and the p-Gröbner basis (Step 3) is the family $\mathcal{G} = \{\{(g, h)\}\}$ because no Svector appears during the computation of the Buchberger algorithm. In this case, Pareto-optimal solutions are obtained as partial remainders of an initial feasible solution (α, α) by (g, h), i.e., the entire set of Pareto-optimal solutions is a subset of $\Gamma = \{\alpha - \lambda g : \lambda \in \mathbb{Z}_+\}$. More explicitly, the set of Pareto-optimal solutions for $MIP_{A,C}(b)$ is the set of minimal elements (with respect to $\prec_C$) of Γ.

ALGORITHM 5: PARETO-OPTIMAL SOLUTIONS COMPUTATION FOR $MIP_{A,C}(b)$.

input : $MIP_{A,C}(b)$.
Step 1. Compute an initial feasible solution α_o for $MIP_{A,C}(b)$ (a solution for the diophantine system of equations $Ax = b$, $x \in \mathbb{Z}^n$).
Step 2. Compute a system of generators for I_A: $\{\{u_i, v_i\} : i = 1, \dots, s\}$ using `setofgenerators(`A`)`.
Step 3. Compute the partial reduced Gröbner basis for $MIP_{A,C}$, $\mathcal{G}_C = \{\mathcal{G}_1, \dots, \mathcal{G}_t\}$ using `pgrobner(` F_1, F_2 `)`, where $F_1 = \{u_i : i = 1, \dots, r\}$ and $F_2 = \{v_i : i = 1, \dots, r\}$.
Step 4. Calculate the set of partial remainders: $R := pRem(\alpha_o, \mathcal{G}_C)$.
output: Pareto-optimal Solutions : R.

To illustrate the above approach, we present an example of MOILP with two objectives where all the computations are done in detail.

Example 3.1.

$$\begin{array}{ll} \min & \{10x + y, x + 10y\} \\ \text{s.t.} & \\ & \begin{array}{rl} 2x + 2y & \geqslant 17, \\ 2y & \leqslant 11, \\ x & \leqslant 10, \\ x, y & \in \mathbb{Z}_+. \end{array} \end{array} \tag{4}$$

Transforming the problem to the standard form results in

$$\begin{array}{ll} \min & \{10x + y + 0z + 0t + 0q, x + 10y + 0z + 0t + 0q\} \\ \text{s.t.} & \\ & \begin{array}{rl} 2x + 2y - z & = 17, \\ 2y + t & = 11, \\ x + q & = 10, \\ x, y, z, t, q & \in \mathbb{Z}_+. \end{array} \end{array} \tag{5}$$

Step 1. Feasible solution for $MIP_{A,C}(b)$: $u = (9, 4, 9, 3, 1)$.
Step 2. Following the steps of Algorithm 3:
1. Basis for $\mathrm{Ker}(A)$: $\mathcal{B} := \{(0, 1, 2, -2, 0), (-1, 0, -2, 0, 1)\}$.
2. LLL-reduced basis for $\mathcal{B}$: $\mathcal{B}_{red} := \mathcal{B} := \{(-1, 0, -2, 0, 1), (-1, 1, 0, -2, 1)\}$.
3. $J_0 := \langle x^{u_+} - x^{u_-} : u \in \mathcal{B}_{red} \rangle = \langle x_5 - x_1x_3^2, x_2x_5 - x_1x_4^2 \rangle$.
4. $J_{i+1} := (J_i : x_i^\infty)$.
   (a) $\widetilde{\mathcal{G}}_0 := \{x_5 - x_1x_3^2, x_2x_5 - x_1x_4^2, x_2x_3^2 - x_4^2\} \Rightarrow J_1 := \langle x_5 - x_1x_3^2, x_2x_5 - x_1x_4^2, x_2x_3^2 - x_4^2 \rangle$.
   (b) $\widetilde{\mathcal{G}}_1 := \{x_5 - x_1x_3^2, x_2x_5 - x_1x_4^2, x_2x_3^2 - x_4^2\} \Rightarrow J_2 := \langle x_5 - x_1x_3^2, x_2x_5 - x_1x_4^2, x_2x_3^2 - x_4^2 \rangle$.
   (c) $\widetilde{\mathcal{G}}_2 := \{x_5 - x_1x_3^2, x_2x_5 - x_1x_4^2, x_2x_3^2 - x_4^2\} \Rightarrow J_3 := \langle x_5 - x_1x_3^2, x_2x_5 - x_1x_4^2, x_2x_3^2 - x_4^2 \rangle$.
   (d) $\widetilde{\mathcal{G}}_3 := \{x_5 - x_1x_3^2, x_2x_5 - x_1x_4^2, x_2x_3^2 - x_4^2\} \Rightarrow J_4 := \langle x_5 - x_1x_3^2, x_2x_5 - x_1x_4^2, x_2x_3^2 - x_4^2 \rangle$.
5. $\Im_A = \langle x_5 - x_1x_3^2, x_2x_5 - x_1x_4^2, x_2x_3^2 - x_4^2, x_1x_3^2 - 1 \rangle \mapsto$
   $I_A = \langle \{\big((1, 0, 0, 0, 1), (0, 1, 0, 2, 0)\big), \big((1, 0, 2, 0, 0), (0, 0, 0, 0, 1)\big), \big((0, 1, 2, 0, 0), (0, 0, 0, 2, 0)\big)\} \rangle$.

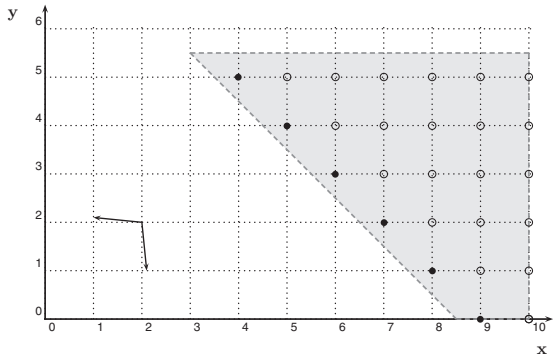


FIG. 4. *Feasible region, Pareto-optimal solutions, and improvement cone for Example* 3.1.

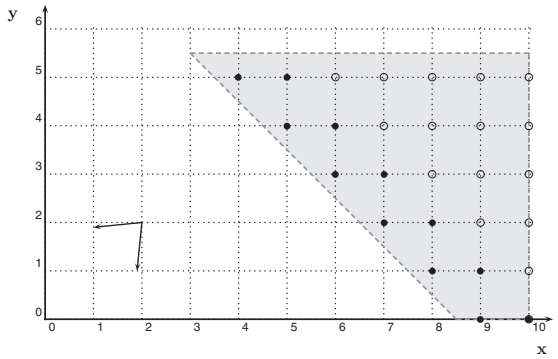


FIG. 5. *Feasible region, Pareto-optimal solutions, and improvement cone for Example* 3.1 *with* $C = [[10, -1], [-1, 10]]$.

Step 3. Computing a p-Gröbner basis for I_A, using order $\prec_C^s$ (Remark 2.2) and following Algorithm 4, we obtain $\mathcal{G}$, whose maximal chains are
$\mathcal{G}_1 = \{\big((0,1,2,0,0),\ (0,0,0,2,0),\ (0,1,2,0,0)\big),\ \big((0,1,0,0,2),\ (2,0,2,2,0),\ (0,1,0,0,2)\big),\ \big((0,1,0,0,1),\ (1,0,0,2,0),\ (0,1,0,0,1)\big)\}$ and
$\mathcal{G}_2 = \{\big((1,0,0,4,0),\ (0,2,2,0,1),\ (1,0,0,4,0)\big),\ \big((1,0,2,0,0),\ (0,0,0,0,1),\ (1,0,2,0,0)\big),\ \big((1,0,0,2,0),\ (0,1,0,0,1),\ (1,0,0,2,0)\big)\}$.

Step 4. Partial remainders. Reducing first by $\mathcal{G}_1$,
$pRem((9,4,9,3,1), \mathcal{G}_1) = \{(9,0,1,11,1)\}$.
Then, reducing each remainder by $\mathcal{G}_2$,
$pRem((9,0,1,11,1), \mathcal{G}_2) = \{(9,0,1,11,1),\ (8,2,3,7,2),\ (7,2,1,9,3),\ (6,3,1,5,4),\ (5,4,1,3,5),\ (4,5,1,1,6)\}$.
The entire set of Pareto-optimal solutions is

$$\{(9,0,1,11,1), (8,1,1,9,2), (7,2,1,7,3), (6,3,1,5,4), (5,4,1,3,5), (4,5,1,1,6)\}.$$

Figure 4 shows the feasible region and the Pareto-optimal solutions of the example above. In addition, we have evaluated the problem with the same feasible region but choosing a cost matrix such that the respective normal vectors of each of the rows in the matrix form an acute angle. Then, nonsupported solutions appear in the set of Pareto-optimal solutions. Figure 5 shows the Pareto-optimal solutions for the same feasible region and $C = \begin{bmatrix} 10 & -1 \\ -1 & 10 \end{bmatrix}$.

TABLE 1
Summary of computational experiments for knapsack problems.

problem	sogt	pgbt	post	tott	|pos|	|maxch|	steps	act_pGB
knap4_2	0.063	249.369	1.265	250.697	11	20	2	164.920
knap4_3	0.063	1002.689	2.012	1004. 704	5	46	2	772.772
knap4_4	0.063	1148.574	2.374	1151.011	16	98	2.4	763.686
knap5_2	0.125	1608.892	0.875	609.892	3	29	2	1187.201
knap5_3	0.125	3500.831	2.035	3503.963	2	30	2.2	2204.123
knap5_4	0.125	3956.534	2.114	3958.773	9	45.4	3	3044.157
knap6_2	0.185	2780.856	2.124	2783.165	18	156	2.4	2241.091
knap6_3	0.185	3869.156	2.018	3871.359	16.4	189	2.4	2790.822
knap6_4	0.185	4598.258	3.006	4601.449	26	298	3.2	3096.466

4. Computational results. A series of computational experiments have been performed in order to evaluate the behavior of the proposed solution method. Programs have been coded in MAPLE 10 and executed in a PC with an Intel Pentium 4 processor at 2.66GHz and 1 GB of RAM. In the implementation of Algorithm 4 to obtain the p-Gröbner basis, the package *poset* for Maple [35] has been used to compute, at each iteration, the maximal chains for the p-Gröbner basis. The implementation has been done in a symbolic programming language, available upon request, in order to make the access easy to both optimizers and algebraic geometers.

The performance of the algorithm was tested on randomly generated instances for knapsack and transportation [29] multiobjective problems for 2, 3, and 4 objectives. For the knapsack problems, 4, 5, and 6 variables, programs have been considered, and, for each group, the coefficients of the constraint were randomly generated in $[0, 20]$, whereas the coefficients of the objective matrices range in $[0, 20]$. Once the constraint vector $(a_1, \ldots, a_n)$ is generated, the right-hand side is fixed as $b = \lceil \frac{1}{2} \sum_{i=1}^{n} a_i \rceil$ to ensure feasibility.

The computational tests for each number of variables have been done in the following way: (1) Generate five constraint vectors and compute the initial system of generators for each of them using Algorithm 3; (2) Generate five random objective matrices for each number of objectives (2, 3, and 4) and compute the corresponding p-Gröbner basis using Algorithm 4; and (3) with $b = \lceil \frac{1}{2} \sum_{i=1}^{n} a_i \rceil$ and for each objective matrix, compute the Pareto-optimal solutions using Algorithm 5.

Table 1 contains a summary of the average results obtained for the considered knapsack multiobjective problems. The second, third, and fourth columns show the average CPU times for each stage in the algorithm: `sogt` is the CPU time for computing the system of generators, `pgbt` is the CPU time for computing a p-Gröbner basis, and `post` is the time for computing a feasible solution and partially reducing it to obtain the set of Pareto-optimal solutions. The fifth column shows the total time for computing the set of Pareto-optimal solutions for the problem. Finally, the sixth and seventh columns show the average number of Pareto-optimal solutions and the number of maximal chains in the p-Gröbner basis for the problem, respectively. The problems have been named as `knapN_O`, where `N` is the number of variables and `O` is the number of objectives. For the transportation problems, instances with 3 origins $\times$ 2 destinations, 3 origins $\times$ 3 destinations, and 4 origins $\times$ 2 destinations have been considered. In this case, for each fixed numbers of origins s and destinations d, the constraint matrix $A \in Z^{(s+d)\times(sd)}$ is fixed. Then, we have generated five instances for each problem of size $s \times d$. Each of these instances is combined with five different right-hand side vectors. The procedure is analogous to the knapsack computational

TABLE 2
Summary of computational experiments for the battery of multiobjective transportation problems.

problem	sogt	pgbt	post	tott	|pos|	|maxch|	steps	act_pGB
tr3x2_2	0.015	11.813	0.000	11.828	5.2	6	2	7.547
tr3x2_3	0.015	7.218	13.108	30.341	12	2.6	2	6.207
tr3x2_4	0.015	6.708	15.791	21.931	6	5	2.2	4.561
tr3x3_2	0.047	1545.916	1.718	1547.681	5	92	2	928.222
tr3x3_3	0.047	3194.333	11.235	3205.615	9	122	2.4	2172.146
tr3x3_4	0.047	3724.657	7.823	3732.527	24	187.4	2.2	2112.287
tr4x2_2	0.046	675.138	2.122	677.306	3.4	35.2	2	398.093
tr4x2_3	0.046	1499.294	6.288	1505.628	5.8	42.4	2.2	119.519
tr4x2_4	0.046	2285.365	7.025	2292.436	12	59	2.2	1654.048

test: a first step where a system of generators is computed, a second one where the p-Gröbner basis is built, and in the last step, the set of Pareto-optimal solutions is computed using partial reductions. Table 2 shows the average CPU times and the average number of Pareto-optimal solutions and maximal chains in the p-Gröbner basis for each problem. The `steps` column shows the average number of steps in the p-Gröbner computation, and `act_pGB` is the average CPU time in the computation of the p-Gröbner basis elapsed since the last element was added to the basis until the end of the process. The problems have been named as `trNxM_O`, where `N` is the number of origins, `M` is the number of destinations, and `O` is the number of objectives. As can be seen in Tables 1 and 2, the overall CPU times are clearly divided into three steps, the most costly being the computation of the p-Gröbner basis. In all of the cases, more than 99% of the total time is spent computing the p-Gröbner basis. Once this structure is computed, obtaining the Pareto-optimal solutions is done very efficiently.

The CPU times and sizes in the different steps of the algorithm are highly sensitive to the number of variables. However, our algorithm is not very sensitive to the number of objectives, since the increment of CPU times with respect to the number of objectives is much smaller than the one with respect to the number of variables.

It is clear that one can not expect fast algorithms for solving MOILP, since all these problems are NP-hard. Nevertheless, our approach provides exact tools that, apart from solving these problems, give insights into the geometric and algebraic nature of the problem.

As mentioned above, using our methodology one can identify the common algebraic structure within any MOILP problem. This connection allows us to improve the efficiency of our algorithm, making use of any advance that improves the computation of Gröbner bases. In fact, any improvements of the standard Gröbner bases theory may have an impact in improving the performance of this algorithm. In particular, one can expect improvements in the efficiency of our algorithm based on the special structure of the integer program (see, for instance, Remark 3.1). In addition, we have to mention another important issue in our methodology. As shown in Theorem 3.2, solving MOILP with the same constraint and objective matrices requires computing only once the p-Gröbner basis. Therefore, once this is done, we can solve different instances varying the right-hand side very quickly.

Finally, we have observed from our computational tests that a significant amount of the time, more than 60% (see column `act_pGB`) for the computation of the p-Gröbner basis is spent checking that no new elements are needed in this structure. This implies that the actual p-Gröbner basis is obtained much earlier than when the final test is finished. A different truncation strategy may be based on the number of steps required to obtain the p-Gröbner basis. According to the exact method, the

algorithm stops once in a step; no new elements are added to the structure. Our tables show that, in most cases, the number of steps is two, actually, only one step is required to generate the entire p-Gröbner basis (see column `steps`). These facts can be used to accelerate the computational times at the price of obtaining only heuristic Pareto-optimal solutions. This idea may be considered an alternative primal heuristic in MOILP and will be the subject of further research.

REFERENCES

[1] K. Aardal, R. Weismantel, and L. Wolsey, *Non-standard approaches to integer programming*, Discrete Appl. Math., 123 (2002), pp. 5–74.

[2] W. Adams and P. Loustaunau, *An Introduction to Gröbner Bases*, Grad. Stud. Math. 3, AMS, Providence, RI, 1994.

[3] M.A. Bazaraa, H.D. Sherali, and C.M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley and Sons, New York, 1993.

[4] R.M. Baer and O. Østerby, *Algorithms over partially ordered sets*, J. BIT Numer. Math., 9 (1969), pp. 97–118.

[5] D. Bertsimas and R. Weismantel, *Optimization Over Integers*, Dynamic Ideas, Belmont, MA, 2005.

[6] B. Buchberger, *An Algorithm for finding the basis elements of the residue class ring of a zero-dimensional polynomial ideal*, J. Symb. Comp., 4 (2005), pp. 475–511.

[7] G. Cantor, *Beiträge zur Begründung der transfiniten Mengenlehre (Zweiter Artikel)*, Math. Ann., 49 (1897), pp. 207–246.

[8] A. Cayley, *A theorem on trees*, Q. J. Math., 23 (1889), pp. 376–378.

[9] V. Chankong and Y.Y. Haimes, *Multiobjective Decision Making Theory and Methodology*. Elsevier Science, New York, 1983.

[10] P. Conti and C. Traverso, *Buchberger algorithm and integer programming*, in Proceedings of the AAECC-9, New Orleans, Lect. Notes Comput. Sci. 539, H. F. Mattson, T. Mora, and T. R. N. Rao, eds., Springer, New York, 1991, pp. 130–139.

[11] D. Cox, J. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, 1st ed., Springer, New York, 1992.

[12] D. Cox, J. Little, and D. O'Shea, *Using Algebraic Geometry*, 1st ed., Springer, New York, 1998.

[13] X. Delorme, X. Gandibleux, and F. Degoutin, *Resolution approché du probleme de set packing bi-objectifs*, in Proceedings of de L'ecole d'Automne de Recherche Operationnelle de Tours (EARO), 2003, pp. 74–80.

[14] F.Y. Edgeworth, *Mathematical Psychics*, P. Keagan, London, 1881.

[15] M. Ehrgott, *Multicriteria Optimization*, Lecture Notes in Econom. Math. Systems 491, Springer, Berlin, 2000.

[16] M. Ehrgott and X. Gandibleux, *A survey and annotated bibliography of multicriteria combinatorial optimization*, OR Spectrum, 22 (2000), pp. 425–460.

[17] M. Ehrgott and X. Gandibleux, eds., *Multiple Criteria Optimization. State of the Art Annotated Bibliographic Surveys*, Kluwer, Boston, 2002.

[18] M. Ehrgott, J. Figueira, and S. Greco, eds., *Multiple Criteria Decision Analysis. State of the Art Surveys*, Springer, New York, 2005.

[19] M. Ehrgott, J. Figueira, and X. Gandibleux, eds., *Multiobjective discrete and combinatorial optimization*, Ann. Oper. Res., 147 (2006), pp. 1–3.

[20] H. Hamacher and G. Ruhe, *On spanning tree problems with multiple objectives*, Ann. Oper. Res., 52 (1994), pp. 209–230.

[21] F. Hausdorff, *Untersuchungen über Ordungtypen, Berichte über die Verhandlungen der königlich sächsischen Gesellschaft der Wissenschaften zu Leipzig*, Matematisch - Physische Klasse, 58 (1906), pp. 106–169.

[22] S. Hoşten and B. Sturmfels, *GRIN: An implementation of Gröbner bases for integer programming*, in Proceedings of the 4th International IPCO Conference, Integer Programming and Combinatorial Optimization, Lect. Notes Comput. Sci. 920, E. Balas and J. Clausen, eds., Springer, Berlin, 1995, pp. 267–276.

[23] S. Hoşten, *Degrees of Gröbner bases of integer programs*, Ph.D. thesis, Cornell University, Ithaca, NY, 1997.

[24] H. Ishibuchi and T. Murata, *A multi-objective genetic local search algorithm and its application to flowshop scheduling*, IEEE Trans. Syst. Man Cybern. C, 28 (1998), pp. 392–403.
[25] N. Jozefowiez, F. Semet, and E-G. Talbi, *A multi-objective evolutionary algorithm for the covering tour problem*, Applications of Multi-Objective Evolutionary Algorithms, C. A. Coello and G. B. Lamont, eds., World Scientific, River Edge, NJ, 2004, pp. 247–267.
[26] M.H. Karwan and B. Villarreal, *Multicriteria dynamic programming with an application to the integer case*, J. Optim. Theory Appl., 31 (1982), pp. 43–69.
[27] A.K. Lenstra, H.W. Lenstra, and L. Lovàsz, *Factoring polynomials with rational coefficients*, Math. Ann., 261 (1982), pp. 515–534.
[28] K. Miettinen, *Nonlinear Multiobjective Optimization*, Kluwer, Boston, 1999.
[29] G.L. Nemhauser and L.A. Wolsey, *Integer and Combinatorial Optimization*, John Wiley and Sons, New York, 1988.
[30] V. Pareto, *Manual d'Economie Politique*, F. Rouge, Lausanne, 1896.
[31] L. Pottier, *Minimal solutions of linear diophantine systems: Bounds and algorithms*, in Proceedings of the Fourth International Conference on Rewriting Techniques and Applications, 1991, Como, Italy, pp. 162–173.
[32] Y. Sawaragi, H. Nakayama, and T. Tanino, *Theory of Multiobjective Optimization*, Academic Press, New York, 1985.
[33] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency*, Springer, New York, 2003.
[34] N. El-Sherbeny, *Resolution of a Vehicle Routing Problem with Multiobjective Simulated Annealing Method*, Ph.D. thesis, Faculte Polytechnique de Mons, Mons, Belgium, 2001.
[35] J.R. Stembridge, *A Maple Package for Posets*, http://www.math.lsa.umich.edu/~jrs.
[36] R.E. Steuer, *Multiple Criteria Optimization: Theory, Computation and Application*, John Wiley and Sons, New York, 1985.
[37] B. Sturmfels, *Gröbner Bases and Convex Polytopes*, AMS, Univ. Lectures Ser. 8, Providence, RI, 1996.
[38] B. Sturmfels, *Solving Systems of Polynomial Equations*, CBMS Reg. Conf. Ser. 97, AMS, Providence, RI, 2002.
[39] B. Sturmfels, *Algebraic Recipes for Integer Programming*, Proc. Sympos. Appl. Math., 61, AMS, Providence. RI, 2004.
[40] B. Sturmfels and R.R. Thomas, *Variation of cost functions in integer programming*, Math. Program., 77 (1997), pp. 357–387.
[41] R.R. Thomas, *A geometric Buchberger algorithm for integer programming*, Math. Oper. Res., 20 (1995), pp. 864–884.
[42] R.R. Thomas, *Applications to integer programming*, in Applications of Computational Algebraic Geometry, D. A. Cox and B. Sturmfels, eds., Proceedings of the 53rd Symposium in Applied Mathematics, AMS, 1997, pp. 119–142.
[43] R.R. Thomas and R. Weismantel, *Truncated Groebner bases for integer programming*, Appl. Algebra Engrg., Comm. Comput., 8 (1997), pp. 241–257.
[44] R. Urbaniak, R. Weismantel, and G.M. Ziegler, *A variant of the Buchberger algorithm for integer programming*, SIAM J. Discrete Math., 10 (1997), pp. 96–108.
[45] P.L. Yu, *Cone convexity, cone extreme points and nondominated solutions in decision problems with multiobjectives*, J. Optim. Theory Appl., 14 (1974), pp. 319–377.
[46] S. Zionts, *A survey of multiple criteria integer programming methods*, Ann. Discrete Math., 5 (1979), pp. 389–398.
[47] S. Zionts and J. Wallenius, *Identifying efficient vectors: Some theory and computational results*, Oper. Res., 23 (1980), pp. 785–793.

SIAM J. DISCRETE MATH.
Vol. 23, No. 2, pp. 596–608

FOURIER SPECTRA OF BINOMIAL APN FUNCTIONS*

CARL BRACKEN†, EIMEAR BYRNE†, NADYA MARKIN†, AND GARY MCGUIRE†

Abstract. In this paper we compute the Fourier spectra of some recently discovered binomial almost perfect nonlinear (APN) functions. One consequence of this is the determination of the nonlinearity of the functions, which measures their resistance to linear cryptanalysis. Another consequence is that certain error-correcting codes related to these functions have the same weight distribution as the 2-error-correcting Bose–Chaudury–Hocquenghem (BCH) code. Furthermore, for field extensions of $\mathbb{F}_2$ of odd degree, our results provide an alternative proof of the APN property of the functions.

Key words. almost perfect nonlinear, APN, Fourier spectrum

AMS subject classifications. 11T23, 11T71, 94B05

DOI. 10.1137/080717079

1. Introduction.

Highly nonlinear functions on finite fields are interesting from the point of view of cryptography as they provide optimum resistance to linear and differential attacks. A function that has the almost perfect nonlinear (APN) (resp., almost bent (AB)) property, as defined below, has optimal resistance to a differential (resp., linear) attack. For more on relations between linear and differential cryptanalysis, see [11].

Highly nonlinear functions are also of interest from the point of view of coding theory. The weight distribution of a certain error-correcting code is equivalent to the Fourier spectrum (including multiplicities) of f. The code having three particular weights is equivalent to the AB property, when n is odd. The minimum distance of the dual code being 5 is equivalent to the APN property holding for f. We give more details on the connections to coding theory in section 2.

For the rest of the paper, let $L = GF(2^n)$, and let L^* denote the set of nonzero elements of L. Let $\mathrm{Tr} : L \to GF(2)$ denote the trace map from L to $GF(2)$.

DEFINITION 1.1. *A function $f : L \to L$ is said to be* almost perfect nonlinear (APN) *if for any $a \in L^*, b \in L$, we have*

$$|\{x \in L : f(x+a) - f(x) = b\}| \leq 2.$$

DEFINITION 1.2. *Given a function $f : L \to L$, the* Fourier transform *of f is the function $\widehat{f} : L \times L^* \to \mathbb{Z}$ given by*

$$\widehat{f}(a,b) = \sum_{x \in L} (-1)^{\mathrm{Tr}(ax+bf(x))}.$$

*Received by the editors February 28, 2008; accepted for publication (in revised form) October 28, 2008; published electronically February 6, 2009.

http://www.siam.org/journals/sidma/23-2/71707.html

†School of Mathematical Sciences, University College Dublin, Dublin 4, Ireland (carlbracken@yahoo.com, ebyrne@ucd.ie, nadyaomarkin@gmail.com, gary.mcguire@ucd.ie). The research of the first author was supported by the Irish Research Council for Science, Engineering and Technology Postdoctoral Fellowship. The research of the second and fourth authors and the Postdoctoral Fellowship of the third author were supported by the Claude Shannon Institute, Science Foundation Ireland grant 06/MI/006.

The *Fourier spectrum* of f is the set of integers

$$\Lambda_f = \{\widehat{f}(a,b) : a, b \in L, b \neq 0\}.$$

The nonlinearity of a function f on a field $L = GF(2^n)$ is defined as

$$NL(f) := 2^{n-1} - \frac{1}{2} \max_{x \in \Lambda_f} |x|.$$

The nonlinearity of a function measures its distance to the set of all affine maps on L. We thus call a function *maximally nonlinear* if its nonlinearity is as large as possible. If n is odd, its nonlinearity is upper-bounded by $2^{n-1} - 2^{\frac{n-1}{2}}$, while for n even an upper bound is $2^{n-1} - 2^{\frac{n}{2}-1}$. For odd n, we say that a function $f : L \longrightarrow L$ is AB when its Fourier spectrum is $\{0, \pm 2^{\frac{n+1}{2}}\}$, in which case it is clear from the upper bound that f is maximally nonlinear. We have the following connection (for odd n) between the AB and APN properties: every AB function on L is also APN [11], and, conversely, if f is quadratic and APN, then f is AB [10]. In particular, quadratic APN functions have optimal resistance to both linear and differential attacks. On the other hand, there appears to be no relation between the nonlinearity $NL(f)$ and the APN property of a function f when n is even. The reader is referred to [8] for a comprehensive survey on APN and AB functions. APN functions were first introduced in Nyberg [14].

Recently, the first nonmonomial families of APN functions have been discovered. Below we list the families of quadratic functions known at this time. We remark that, in a sense to be qualified in the next section (namely, Carlet–Charpin–Zinoviev (CCZ) equivalence [10]), these families are all pairwise inequivalent.

(1)

$$f(x) = x^{2^s+1} + \alpha x^{2^{ik}+2^{mk+s}},$$

where $n = 3k$, $(k,3) = (s,3k) = 1$, $k \geq 3$, $i \equiv sk \bmod 3$, $m \equiv -i \mod 3$, $\alpha = t^{2^k-1}$, and t is primitive (see Budaghyan et al. [6], [4]).

(2)

$$f(x) = x^{2^s+1} + \alpha x^{2^{ik}+2^{mk+s}},$$

where $n = 4k$, $(k,2) = (s,2k) = 1$, $k \geq 3$, $i \equiv sk \bmod 4$, $m = 4-i$, $\alpha = t^{2^k-1}$, and t is primitive (see Budaghyan, Carlet, and Leander [5]). This family generalizes an example found for $n = 12$ by Edel, Kyureghyan, and Pott [13].

(3)

$$f(x) = \alpha x^{2^s+1} + \alpha^{2^k} x^{2^{k+s}+2^k} + \beta x^{2^k+1} + \sum_{i=1}^{k-1} \gamma_i x^{2^{k+i}+2^i},$$

where $n = 2k$, α and β are primitive elements of $GF(2^n)$, $\gamma_i \in GF(2^k)$ for each i, $(k,s) = 1$, k is odd, and s is odd (see Bracken et al. [1]).

(4)

$$f(x) = x^3 + Tr(x^9),$$

over $GF(2^n)$, for any n (see Budaghyan, Carlet, and Leander [7]).

(5)

$$f(x) = ux^{2^{-k}+2^{k+s}} + u^{2^k}x^{2^s+1} + vx^{2^{k+s}+2^s},$$

where $n = 3k$, u is primitive, $v \in GF(2^k), (s, 3k) = 1$, $(3, k) = 1$, and 3 divides $k + s$ (see Bracken et al. [1]).

In this paper we calculate the Fourier spectra of families (1) and (2). The determination of the Fourier spectra of families (3) and (4) has been given in [2] and [3], respectively, using other methods. The Fourier spectrum for family (5) has not yet been found and is an open problem. We will show here that the Fourier spectra of the functions (1) and (2) are 5-valued for fields of even degree and 3-valued for fields of odd degree. In this sense they resemble the Gold functions x^{2^d+1}, $(d, n) = 1$. For fields of odd degree, our result provides another proof of the APN property. This does not hold for fields of even degree; as we stated earlier, there appears to be no relation between the Fourier spectrum and the APN property for fields of even degree. Thus, the fact that f has a 5-valued Fourier spectrum for fields of even degree does not follow from the fact that f is a quadratic APN function. Indeed, there is one example known (due to Dillon [12]) of a quadratic APN function on a field of even degree whose Fourier spectrum is more than 5-valued; if u is primitive in $GF(2^6)$, then

$$g(x) = x^3 + u^{11}x^5 + u^{13}x^9 + x^{17} + u^{11}x^{33} + x^{48}$$

is a quadratic APN function on $GF(2^6)$ whose Fourier transform takes seven distinct values.

The layout of this paper is as follows. In section 2 we review the connections between APN functions, nonlinearity, and coding theory. Section 3 gives the proof of the Fourier spectrum for family (1), and section 4 gives the proof for family (2). In section 5 we simply state for completeness the results from other papers on families (3) and (4), and section 6 has some open problems for further work.

2. Preliminaries on coding theory. Fix a basis of L over $\mathbb{F}_2$. For each element $x \in L$ we write $\mathbf{x} = (x_1, \ldots, x_n)$ to denote the vector of coefficients of x with respect to this basis. Given a map $f : L \longrightarrow L$, we write $f(\mathbf{x})$ to denote the representation of $f(x) \in L$ as a vector in $\mathbb{F}_2^n$, and we consider the $2n \times (2^n - 1)$ binary matrix

$$A_f = \begin{bmatrix} \cdots & \mathbf{x} & \cdots \\ \cdots & f(\mathbf{x}) & \cdots \end{bmatrix},$$

where the columns are ordered with respect to some ordering of the nonzero elements of L.

The function f is APN if and only if the binary error-correcting code of length $2^n - 1$ with A_f as parity check matrix has a minimum of distance 5. This is because codewords of weight 4 correspond to solutions of

$$a + b + c + d = 0,$$
$$f(a) + f(b) + f(c) + f(d) = 0,$$

and this system has no nontrivial solutions if and only if f is APN. We refer the reader to [9] for more on the connection between coding theory and APN functions.

The dual code has A_f as a generator matrix. The weights w in this code correspond to values V in the Fourier spectrum of f via $V = n - 2w$. Thus, when we compute the Fourier spectrum of an APN function, as we do in this paper, we are computing the weights occurring in the code.

Suppose f is APN. Let C_f denote the code with generator matrix A_f. Let a_w denote the number of times the weight w occurs in C_f. Let b_j denote the number of codewords of weight j in $C_f^{\perp}$. If there are five or fewer weights in C_f, the MacWilliams (or Pless) identities yield five independent equations, $b_0 = 1$, $b_1 = b_2 = b_3 = b_4 = 0$, for the unknowns a_V, which can be solved uniquely. Thus the distribution of values is determined for an APN function whenever there are five or fewer values in its Fourier spectrum. In particular, if $\Lambda_f \subseteq \{0, \pm 2^{\frac{n}{2}}, \pm 2^{\frac{n+2}{2}}\}$ for even n, or $\Lambda_f \subseteq \{0, \pm 2^{\frac{n+1}{2}}\}$ for odd n, then the distribution is completely determined. This is indeed the case for the functions studied in this paper. Solving for the distribution in this case must yield the same values and distribution as the double-error-correcting BCH code, which corresponds to the APN function x^3. This function has $\Lambda_f = \{0, \pm 2^{\frac{n}{2}}, \pm 2^{\frac{n+2}{2}}\}$ for even n, and $\Lambda_f = \{0, \pm 2^{\frac{n+1}{2}}\}$ for odd n.

Consider the extended code of $C_f^{\perp}$, which has parity check matrix

$$P_f = \begin{bmatrix} 1 \cdots & 1 & 1 & 1 \\ \cdots & x & \cdots & 0 \\ \cdots & f(x) & \cdots & 0 \end{bmatrix}.$$

Two functions f and g are said to be CCZ equivalent [10] if and only if the codes with parity check matrices P_f and P_g are equivalent (as binary codes). This is not the original definition of CCZ equivalence, but it is an equivalent definition, as was shown in [1].

The new APN functions presented in the introduction are known to be pairwise CCZ inequivalent. One consequence of the results in this paper is that further invariants (beyond the code weight distribution) are needed to show that families (1)–(4) are inequivalent.

3. Family (1), binomials over $GF(2^{3k})$. We will make good use of the following standard result from Galois theory, which allows us to bound the number of solutions of a linearized polynomial. We include a proof for the convenience of the reader.

LEMMA 3.1. *Let F be a field, and let K, H be finite Galois extensions of F of degrees n and s, respectively, whose intersection is F. Let $M = KH$ be the compositum of K and H. Let $k_1, \ldots, k_t$ be F-linearly independent elements of K. Then $k_1, \ldots, k_t$ are H-linearly independent when regarded as elements of M.*

Proof. Since K and H are Galois extensions of F and $K \bigcap H = F$, we have $[M : H] = [K : F] = n$. Let $\{k_1, \ldots, k_n\}$ be an F-basis of K as a vector space over F, and let $\{h_1, \ldots, h_s\}$ be an F-basis of H as a vector space over F. Then the set $\{k_i \cdot h_j \mid 1 \le i \le n, 1 \le j \le s\}$ generates M as a vector space over F. It is clear that the set $\{k_1 \cdot h_1, \ldots, k_n \cdot h_1\}$ generates M as a vector space over the field H. Without loss of generality we can assume that $h_1 = 1$. Since $[M : H] = n$, we conclude that $\{k_1, \ldots, k_n\}$ is indeed a basis of M over H.

Let $\{k_1, \ldots, k_t\}$ be a set of F-linearly independent elements of K. We can extend this set to a basis $\{k_1, \ldots, k_t, \ldots, k_n\}$. Since this set forms an H-basis of M, its subset $\{k_1, \ldots, k_t\}$ is a fortiori linearly independent over H. ∎

Note that for Galois extensions K, H in the lemma above, $(s,n)=1$ implies that $K \bigcap H = F$, and in the case when the fields K, H, F are finite, we have $(s,n)=1$ if and only if $K \bigcap H = F$.

We will apply Lemma 3.1 to obtain an estimate on the number of zeroes of linearized polynomials. This idea has appeared in the literature before, such as in Trachtenberg's Ph.D. thesis [15].

COROLLARY 3.2. *Let s be an integer satisfying $(s,n)=1$, and let $f(x)=\sum_{i=0}^{d} r_i x^{2^{si}}$ be a polynomial in $L[x]$. Then $f(x)$ has at most 2^d zeroes in L.*

Proof. Let V denote the set of zeroes of $f(x)$ in L. We may assume that $V \neq \{0\}$. Since $f(x)$ is a linearized polynomial, V is a vector space over $GF(2)$ of finite dimension v for some positive integer v. Let $V' \subset GF(2^{sn})$ denote the vector space generated by the elements of V over the field $GF(2^s)$. Since $(s,n)=1$, we have $L \bigcap GF(2^s) = GF(2)$, and by Lemma 3.1, V' is a v-dimensional vector space over $GF(2^s)$. Furthermore, for all $c \in GF(2^s)$ and $w \in GF(2^{sn})$ we have $f(cw)=cf(w)$. Therefore all the elements of V' are also zeroes of $f(x)$. Since the dimension of V over $GF(2)$ is v, the size of V' is 2^{sv}, and it follows that there are at least 2^{sv} zeroes of $f(x)$ in $GF(2^{sn})$. On the other hand, polynomials of degree 2^{ds} can have at most 2^{ds} solutions. We conclude that $v \leq d$. ∎

THEOREM 3.3. *Let*

$$f(x) = x^{2^s+1} + \alpha x^{2^{ik}+2^{mk+s}},$$

where $n=3k$, $(k,3)=(s,3k)=1$, $k \geq 3$, $i \equiv sk \bmod 3$, $m \equiv -i \bmod 3$, $\alpha = t^{2^k-1}$, and t is primitive in L.

The Fourier spectrum of $f(x)$ is $\{0, \pm 2^{\frac{n+1}{2}}\}$ when n is odd and $\{0, \pm 2^{\frac{n}{2}}, \pm 2^{\frac{n+2}{2}}\}$ when n is even.

Proof. By the restrictions on i, s, k, there are two possibilities for our function $f(x)$:

$$f_1(x) = x^{2^s+1} + \alpha x^{2^{-k}+2^{k+s}}, \qquad sk \equiv -1 \mod 3$$

and

$$f_2(x) = x^{2^s+1} + \alpha x^{2^k+2^{-k+s}}, \qquad sk \equiv 1 \mod 3.$$

Let us consider the first case, when $f = f_1$. By definition, the Fourier spectrum of f is

$$f^W(a,b) = \sum_u (-1)^{\mathrm{Tr}(ax+bf(x))}.$$

Squaring gives

$$\begin{aligned} f^W(a,b)^2 &= \sum_{x \in L} \sum_{y \in L} (-1)^{\mathrm{Tr}(ax+bf(x)+ay+bf(y))} \\ &= \sum_{x \in L} \sum_{u \in L} (-1)^{\mathrm{Tr}(ax+bf(x)+a(x+u)+bf(x+u))} \end{aligned}$$

from the substitution $y = x+u$.

This becomes

$$f^W(a,b)^2 = \sum_u (-1)^{\mathrm{Tr}(au+bu^{2^s+1}+b\alpha u^{2^{-k}+2^{k+s}})} \sum_x (-1)^{\mathrm{Tr}(xL_b(u))},$$

where $L_b(u) := bu^{2^s} + (bu)^{2^{-s}} + (b\alpha)^{2^k} u^{2^{-k+s}} + (b\alpha)^{2^{-k-s}} u^{2^{k-s}}$.

Using the fact that $\sum_x (-1)^{\mathrm{Tr}(cx)}$ is 0 when $c \neq 0$ and 2^n otherwise, we obtain

$$f^W(a,b)^2 = 2^n \sum_{u \in K} (-1)^{\mathrm{Tr}(au+bu^{2^s+1}+b\alpha u^{2^{-k}+2^{k+s}})},$$

where K denotes the kernel of $L_b(u)$. If the size of the kernel is at most 4, then clearly

$$0 \leq \sum_{u \in K} (-1)^{\mathrm{Tr}(au+bu^{2^s+1}+b\alpha u^{2^{-k}+2^{k+s}})} \leq 4.$$

Since $f^W(a,b)$ is an integer, this sum can be only 0, 2, or 4 if n is even, and 1 or 3 if n is odd. The set of permissible values of $f^W(a,b)$ is then

$$f^W(a,b) \in \begin{cases} \{0, \pm 2^{\frac{n+1}{2}}\}, & 2 \nmid n, \\ \{0, \pm 2^{\frac{n}{2}}, \pm 2^{\frac{n+2}{2}}\}, & 2 \mid n. \end{cases}$$

We must now demonstrate that $|K| \leq 4$, which is sufficient to complete the proof.

Note that since α is a (2^k-1)nd power, we have $\alpha^{2^{2k}+2^k+1} = 1$. Now suppose that $L_b(u) = 0$. Then we have the following equations:

$$(b\alpha)^{-2^k} L_b(u) + b^{1-2^k-2^{-k}} \alpha L_b(u)^{2^k} + b^{-2^{-k}} L_b(u)^{2^{2k}} = 0,$$

$$b^{-2^{-s}} L_b(u) + b^{2^{-k-s}-2^{k-s}-2^{-s}} \alpha^{2^{-k-s}} L_b(u)^{2^k} + b^{-2^{k-s}} \alpha^{-2^{k-s}} L_b(u)^{2^{-k}} = 0.$$

Substituting the definition of $L_b(u)$ into the above equations and gathering the terms gives

$$c_1 u^{2^{-s}} + c_2 u^{2^{k-s}} + c_3 u^{2^{-k-s}} = 0, \tag{3.1}$$

$$d_1 u^{2^s} + d_2 u^{2^{k+s}} + d_3 u^{2^{-k+s}} = 0, \tag{3.2}$$

where the coefficients c_i, d_j are defined by

$$\begin{aligned}
c_1 &= b^{2^{-s}-2^k} \alpha^{-2^k} + b^{2^{k-s}-2^{-k}} \alpha^{2^{k-s}}, \\
c_2 &= (b\alpha)^{2^{-k-s}-2^k} + b^{2^{k-s}+1-2^{-k}-2^k} \alpha, \\
c_3 &= b^{2^{-s}+1-2^k-2^{-k}} \alpha^{2^{-s}+1} + b^{2^{-k-s}-2^{-k}}, \\
d_1 &= b^{1-2^{-s}} + b^{2^{-k-s}+2^{-k}-2^{-s}-2^{k-s}} \alpha^{2^{-k-s}+2^{-k}}, \\
d_2 &= b^{2^{-k-s}+2^k-2^{-s}-2^{k-s}} \alpha^{2^{-k-s}} + b^{1-2^{k-s}} \alpha^{2^{-k-s}+2^{-s}+1}, \\
d_3 &= b^{2^k-2^{-s}} \alpha^{2^k} + b^{2^{-k}-2^{k-s}} \alpha^{2^{-k-s}+2^{-s}}.
\end{aligned}$$

First we demonstrate that the coefficients c_i, d_j in (3.1) and (3.2) do not vanish. Suppose that $c_1 = 0$. We then have

$$\alpha^{2^{k-s}+2^k} = b^{-2^{k-s}+2^{-k}+2^{-s}-2^k},$$

and taking the 2^{-k}nd power of both sides yields

$$\alpha^{2^{-s}+1} = b^{(2^{k+s}-1)(2^{-s}-2^{-k-s})}.$$

Let $\alpha = t^{2^k-1}$, where t is primitive in $GF(2^{3k})$. Substituting t into the previous equation and rearranging gives

$$t^{2^{k-s}-1} = t^{2^{-s}(1-2^{k+s})} b^{(2^{k+s}-1)(2^{-s}-2^{-k-s})}.$$

The multiplicative order of 2 modulo 7 is equal to 3; therefore for any r we have that 7 divides $2^r - 1$ if and only if r is divisible by 3. Since $sk \equiv -1$, $3 \nmid k - s$, and we conclude that $7 \nmid 2^{k-s} - 1$. Therefore the left-hand side is not a seventh power, while the right-hand side is. We conclude that the coefficient of $u^{2^{-s}}$ in (3.1) is not 0 and use the same type of argument to conclude that all the coefficients in (3.1) are nonzero. A similar argument holds for (3.2).

We will next combine (3.1) and (3.2) to obtain an equation of the form

$$Au + Bu^{2^k} = 0.$$

Raise (3.1) to the power of 2^s, (3.2) to the power of 2^{-s}, and combine the two expressions, cancelling the terms in $u^{2^{-k}}$, to obtain

$$Au + Bu^{2^k} = 0, \tag{3.3}$$

where $A = (\frac{c_1}{c_3})^{2^s} + (\frac{d_1}{d_3})^{2^{-s}}$ and $B = (\frac{c_2}{c_3})^{2^s} + (\frac{d_2}{d_3})^{2^{-s}}$.

For now assume that both A, B are nonzero. We obtain the following equalities by applying the appropriate powers of the Frobenius automorphism to (3.3):

$$u^{2^{-k+s}} = A^{-2^{-k+s}} B^{2^{-k+s}} u^{2^s},$$

$$u^{2^{k-s}} = B^{-2^{-s}} A^{2^{-s}} u^{2^{-s}}.$$

Substituting the two identities above into our expression for $L_b(u) = 0$ gives

$$(b + (b\alpha)^{2^k} A^{-2^{-k+s}} B^{2^{-k+s}}) u^{2^s} + (b^{2^{-s}} + (b\alpha)^{2^{-k-s}} B^{-2^{-s}} A^{2^{-s}}) u^{2^{-s}} = 0. \tag{3.4}$$

Raising this equation to the power of 2^s gives a polynomial of degree 2^{2s} which is $GF(2^s)$ linear. By Corollary 3.2, the dimension of the kernel of this polynomial over $GF(2)$ is at most 2, unless the left-hand side of (3.4) is identically 0. We conclude that $rs \leq 2s$, and hence $r \leq 2$. It therefore remains to show that the polynomial in (3.4) is not identically 0. Assuming that both coefficients are zero, we get

$$Ab^{2^{k-s}} + (b\alpha)^{2^{-k-s}} B = 0,$$
$$Bb + (b\alpha)^{2^{-k}} A = 0.$$

We combine the equations above to obtain

$$Bb + (b\alpha)^{2^{-k}} b^{2^{-k-s}-2^{k-s}} \alpha^{2^{-k-s}} B = 0.$$

So we have $b^{1-2^{-k}+2^{k-s}-2^{-k-s}} = \alpha^{2^{-k-s}+2^{-k}}$. Substituting α with t^{2^k-1}, rearranging, and factoring the powers gives

$$b^{(2^{k+s}-1)(1-2^{-k})} t^{1-2^{k+s}} = t^{2^s(2^{k-s}-1)}.$$

Here we observe that only the left-hand side of the above equation is a seventh power, thus obtaining the desired contradiction. We conclude that the size of the kernel K is less than 4. This finishes the argument.

It finally remains to show that the coefficients A, B are nonzero. Setting A to 0 gives rise to the equation

$$\alpha^{2^{k-2s}+2^{k+s}} = \left(\frac{b^{1-2^{-k+s}} + (b\alpha)^{2^k+2^{k+s}}}{b^{2^{-s}+2^{k-2s}} + b^{2^{-k}+2^{-k-s}} \alpha^{2^{-k-2s}+2^{-k-s}}} \right) \left(\frac{(b\alpha)^{2^{k-s}+2^{k-2s}} + b^{2^{-k-s}-2^{-k-2s}}}{(b\alpha)^{2^s+1} + b^{2^{-k}+2^{k+s}}} \right).$$

Substituting α with t^{2^k-1} and rearranging gives the equation

$$t^{2^{k-2s+1}(2^k-1)} = t^{2^{k-2s}-2^{-k-2s}(2^{3s}-1)} R^{2^{2k+2s}-1} T^{1-2^{2k+2s}},$$

where

$$R = b^{2^{-s}+2^{k-2s}} + b^{2^{-k}+2^{-k-s}} \alpha^{2^{-k-2s}+2^{-k-s}}$$

and

$$T = \left((b\alpha)^{2^{k-s}+2^{k-2s}} + b^{2^{-k-s}-2^{-k-2s}} \right).$$

Reducing the powers of 2 modulo 3 shows that the right-hand side of the equation above is a seventh power, while the left-hand side is not. We conclude that $A \neq 0$.

Suppose $B = 0$; then the only solution of (3.3) is $u = 0$. We can therefore assume that both A and B are nonzero.

This completes the proof of the theorem for the case when $f = f_1$. When $f = f_2$, a similar proof applies. We interchange k and $-k$ in all equations and use the fact that in this case 3 divides $k - s$. ∎

4. Family (2), binomials over $GF(2^{4k})$. We now compute the Fourier spectrum for family (2).

THEOREM 4.1. *Let $L = GF(2^n)$ and $f(x) = x^{2^s+1} + \alpha x^{2^{ik}+2^{mk+s}}$, where $n = 4k$, $(k, 2) = (s, 2k) = 1$, $k \geq 3$, $i \equiv sk \bmod 4$, $m = 4 - i$, $\alpha = t^{2^k-1}$, and t is primitive. Then f has Fourier spectrum $\{0, \pm 2^{n/2}, \pm 2^{\frac{n+2}{2}}\}$.*

Proof. As discussed in the proof of the previous theorem, since f is APN, it suffices to demonstrate that the equation

$$L_b(u) = bu^{2^s} + (bu)^{2^{-s}} + (b\alpha)^{2^{mk}} u^{2^{2k+s}} + (b\alpha)^{2^{ik-s}} u^{2^{2k-s}} = 0$$

has at most four solutions for all nonzero b in L.

Since s, k are chosen to be odd, $sk \equiv \pm 1 \bmod 4$. Therefore there are two possibilities for our function $f(x)$:

$$f_1(x) = x^{2^s+1} + \alpha x^{2^{-k}+2^{k+s}}, \qquad sk \equiv -1 \mod 4$$

and

$$f_2(x) = x^{2^s+1} + \alpha x^{2^k+2^{-k+s}}, \qquad sk \equiv 1 \mod 4.$$

Let us consider the first case, when $f = f_1$, so that

$$L_b(u) = bu^{2^s} + (bu)^{2^{-s}} + (b\alpha)^{2^k} u^{2^{2k+s}} + (b\alpha)^{2^{-k-s}} u^{2^{2k-s}}.$$

All the solutions of $L_b(u) = 0$ are also solutions of the equation

$$b^{-2^{2k}} L_b(u)^{2^{2k}} + (b\alpha)^{-2^k} L_b(u) = 0.$$

Taking this sum results in an elimination of the term in $u^{2^{2k+s}}$. Now multiply by $b^{2^k+2^{2k}}\alpha^{2^k}$ to obtain

$$(b^{2^{2k}+1} + (b\alpha)^{2^k+2^{-k}})u^{2^s} + (b^{2^{2k}+2^{-s}} + (b\alpha)^{2^k+2^{k-s}})u^{2^{-s}}$$

$$+ (b^{2^k+2^{2k-s}}\alpha^{2^{2k}} + b^{2^{2k}+2^{-k-s}}\alpha^{2^{-k-s}})u^{2^{2k-s}} = 0. \tag{4.1}$$

We also compute $b^{-2^{-s+2k}} L_b(u)^{2^{2k}} + (b\alpha)^{-2^{-k-s}} L_b(u) = 0$ to obtain

$$(b^{2^{2k-s}} + (b\alpha)^{2^{-k-s}+2^{-k}})u^{2^s} + (b^{2^{2k-s}+2^{-s}} + (b\alpha)^{2^{k-s}+2^{-k-s}})u^{2^{-s}}$$

$$+ (b^{2^{2k-s}+2^k}\alpha^{2^k} + b^{2^{2k}+2^{-k-s}}\alpha^{2^{-k-s}})u^{2^{2k+s}} = 0. \tag{4.2}$$

Writing (4.2) as

$$cu^{2^s} + du^{2^{-s}} + eu^{2^{2k+s}} = 0, \tag{4.3}$$

we see that (4.1) becomes

$$d^{2^s} u^{2^s} + c^{2^{2k}} u^{2^{-s}} + eu^{2^{2k-s}} = 0. \tag{4.4}$$

We combine (4.3) and (4.4) to cancel the third term from each expression. This yields the equation

$$G(u) := (e^{2^s} c^{2^{-s}} + e^{2^{-s}} c^{2^{2k+s}})u + e^{2^s} d^{2^{-s}} u^{2^{-2s}} + e^{2^{-s}} d^{2^{2s}} u^{2^{2s}} = 0. \tag{4.5}$$

Now for some nonzero v in the kernel of $G(u)$, we consider the equation

$$G_v(u) := uG(u) + vG(v) + (u+v)G(u+v) = 0. \tag{4.6}$$

Substituting gives

$$e^{2^s}d^{2^{-s}}(u^{2^{-2s}}v+v^{2^{-2s}}u)+e^{2^{-s}}d^{2^{2s}}(u^{2^{2s}}v+v^{2^{2s}}u)=0. \tag{4.7}$$

Note that $\ker(G(u))$ is contained in $\ker(G_v(u))$.

We now show that $L_b(u)=0$ has at most four solutions. This will be done in five steps as follows, which completes the proof:

(i) We show that $d\neq 0$ implies that d^{2^s-1} is not a cube.
Recall that $d=b^{2^{2k-s}+2^{-s}}+b^{2^{k-s}+2^{-k-s}}t^{2^{2k-s}+2^{-s}-2^{k-s}-2^{-k-s}}$. This implies that

$$d^{2^s-1}=t^{-2^{-s-k}(2^{2k}+1)(2^s-1)}A^{2^s-1},$$

where $A=b^{2^{2k-s}+2^{-s}}t^{2^{k-s}+2^{-k-s}}+b^{2^{k-s}+2^{-k-s}}t^{2^{2k-s}+2^{-s}}$. As $A=A^{2^k}$, we have $A\in GF(2^k)$. Furthermore, as k is odd, all elements of $GF(2^k)$ are cubes. We conclude that A^{2^s-1} is a cube. Now if d^{2^s-1} is a cube, then so is $t^{(2^{2k}+1)(2^s-1)}$. But this is impossible, as $(2^{2k}+1)(2^s-1)$ is not divisible by 3 and t is primitive.

(ii) We show that if $c,d,e\neq 0$ and d^{2^s-1} is not a cube, then (4.7) has at most four solutions.
Assume that the coefficients e,c,d are nonzero and that d^{2^s-1} is not a cube. Now $u^{2^{2s}}v+v^{2^{2s}}u=0$ if and only if $uv^{-1}\in GF(4)\cap GF(2^{2s})=GF(4)$, since $(s,2k)=1$. Therefore we have exactly four solutions in u, namely, $u=vw$ for each $w\in GF(4)$. If, on the other hand, $u^{2^{2s}}v+v^{2^{2s}}u\neq 0$, we can rearrange (4.7) to obtain

$$d^{2^s-1}=(u^{2^{-2s}}v+v^{2^{-2s}}u)^{2^{2s}-1}e^{2^{-s}-2^s}d^{2^{2s}-1}.$$

Using the fact that 3 divides 2^r-1 if and only if r is even, we see that the right-hand side of this expression is a cube, while the left-hand side is not. Thus, the kernel of L_b has at most four elements.

(iii) We demonstrate that $e\neq 0$.
For the sake of contradiction suppose that $e=0$. Then we have

$$b^{2^{2k-s}+2^k-2^{2k}-2^{-k-s}}t^{2^{2k-s}-2^{-s}-2^k+2^{-k-s}}=1,$$

and hence

$$(bt^{-1})^{2^{2k-s}+2^k-2^{2k}-2^{-k-s}}t^{2^{2k-s}-2^{-s}}=1.$$

Further rearrangement gives

$$(bt^{-1})^{(1-2^k)(2^{2k-s}+2^k)}=t^{2^{-s}(1-2^{2k})}. \tag{4.8}$$

As 4 divides $k+s$, $2^{k+s}\equiv 1 \bmod 5$. Also $2^{2k}+1\equiv 0 \bmod 5$ for any odd k. Therefore 5 divides $2^{k+s}+2^{2k}$, and hence 5 divides 2^k+2^{2k-s}. The left-hand side of (4.8) is a fifth power, while the right-hand side is not because t is primitive and $2^{-s}(1-2^{2k})$ is not a multiple of 5. We conclude that $e\neq 0$.

(iv) We next rule out the case $c=0$.
Suppose $c=0$. Then we have

$$b^{2^{2k-s}+1-2^{-k-s}-2^{-k}}t^{2^{-k-s}+2^{-k}-2^{-s}-1}=1,$$

from which we derive

$$(bt^{-1})^{(2^k-1)(2^{-k}-2^{2k-s})} = t^{2^{-s}(2^{2k}-1)}.$$

By similar observations as before we can demonstrate that only the left-hand side of the expression above is a fifth power. This gives us the desired contradiction, and we conclude that $c \neq 0$.

(v) We show that if $d = 0$, then $L_b(u) = 0$ has at most four solutions. Suppose that $d = 0$. Then (4.4) becomes

$$c^{2^{2k}}u^{2^{-s}} + eu^{2^{2k-s}} = 0. \tag{4.9}$$

Let $H(u) := c^{2^{-s}}u + e^{2^{-s}}u^{2^{2k}}$, so that the solutions to (4.9) make up the kernel of $H(u)$. For some $v \neq 0$ in the kernel of $H(u)$, consider the equation

$$H_v(u) := uH(u) + vH(v) + (u+v)H(u+v) = 0.$$

This yields $H_v(u) = e^{2^{-s}}(u^{2^{2k}}v + v^{2^{2k}}u) = 0$, from which we obtain $u^{2^{2k}} = v^{2^{2k-1}}u$. Applying this relation to $L_b(u) = 0$ gives us the equation

$$L_b(u) = (b + (b\alpha)^{2^k}v^{2^{2k+s}-2^s})u^{2^s} + (b^{2^{-s}} + (b\alpha)^{2^{-k-s}}v^{2^{2k-s}-2^{-s}})u^{2^{-s}} = 0.$$

If both coefficients in the above expression are nonzero, then, by Corollary 3.2, it has at most four solutions. If exactly one of the coefficients is 0, then $u = 0$ is the unique solution. If both coefficients vanish, then taking the first to the 2^{-s}nd power and the second to the 2^snd power gives the equations

$$b^{2^{-s}} + (b\alpha)^{2^{k-s}}v^{2^{2k}-1} = 0$$

and

$$b + (b\alpha)^{2^{-k}}v^{2^{2k}-1} = 0,$$

respectively, from which we derive

$$v^{2^{2k}-1} = b^{2^{-k}-2^{k-s}}\alpha^{-2^{k-s}} = b^{1-2^{-k}}\alpha^{-2^{-k}},$$

which implies that $e = 0$, a previously established contradiction.

This completes the proof of the theorem for the case when $f = f_1$. When $f = f_2$, a near identical proof applies. We simply interchange k and $-k$ in all equations and use the fact that in this case 5 divides $2^{k-s} - 1$ to achieve the required contradictions concerning fifth powers. □

5. Families (3) and (4). For proofs of the following theorems, which compute the Fourier spectra of families (3) and (4), see [2] and [3], respectively. We state the results here for completeness.

THEOREM 5.1. *Let $n = 2k$ and let*

$$f(x) = \alpha x^{2^s+1} + \alpha^{2^k}x^{2^{k+s}+2^k} + \beta x^{2^k+1} + \sum_{i=1}^{k-1}\gamma_i x^{2^{k+i}+2^i},$$

where α and β are primitive elements of L, and $\gamma_i \in GF(2^k)$ for each i and $(k,s) = 1$. Then the Fourier spectrum of $f(x)$ is $\{0, \pm 2^{\frac{n}{2}}, \pm 2^{\frac{n+2}{2}}\}$.

THEOREM 5.2. *Let*

$$f(x) = x^3 + \mathrm{Tr}(x^9)$$

on L. Then the Fourier spectrum of $f(x)$ is $\{0, \pm 2^{\frac{n+1}{2}}\}$ when n is odd and is $\{0, \pm 2^{\frac{n}{2}}, \pm 2^{\frac{n+2}{2}}\}$ when n is even.

6. Closing remarks and open problems. For each of the above quadratic APN functions considered, the Fourier spectrum turned out to be the same as the Gold functions. The example of Dillon on $GF(2^6)$ cited in the introduction of this paper is the only known example of a quadratic APN function that does not have this spectrum. This means that the dual code of this function (as defined in section 2) has the same minimum distance as the double error-correcting BCH code (the dual code corresponding to the function x^3), but has a different weight distribution.

Open problem 1. Find other examples of quadratic APN functions for even n that do not have the same Fourier spectrum as the Gold function x^3.

In [1] the following trinomial function (family (5) in the introduction) over $GF(2^{3k})$ was shown to be APN:

$$f(x) = ux^{2^{-k}+2^{k+s}} + u^{2^k}x^{2^s+1} + vx^{2^{k+s}+2^s},$$

where u is primitive, $v \in GF(2^k), (s, 3k) = 1, (3, k) = 1$, and 3 divides $k + s$.

Open problem 2. Determine the Fourier spectrum of the above APN function.

Acknowledgments. The authors would like to thank the anonymous referees, whose helpful comments greatly improved the presentation of this paper.

REFERENCES

[1] C. Bracken, E. Byrne, N. Markin, and G. McGuire, *New families of quadratic almost perfect nonlinear trinomials and multinomials*, Finite Fields Appl., 14 (2008), pp. 703–714.

[2] C. Bracken, E. Byrne, N. Markin, and G. McGuire, *Determining the nonlinearity of a new family of APN functions*, in Applied Algebra, Algebraic Algorithms and Error Correcting Codes, Lecture Notes in Comput. Sci. 4851, Springer-Verlag, New York, 2007, pp. 72–79.

[3] C. Bracken, E. Byrne, N. Markin, and G. McGuire, *On the Walsh spectrum of a new APN function*, in Cryptography and Coding, Lecture Notes in Comput. Sci. 4887, Springer-Verlag, New York, 2007, pp. 92–98.

[4] L. Budaghyan, C. Carlet, and G. Leander, *A Class of Quadratic APN Binomials Inequivalent to Power Functions*, preprint, 2006. Available online at http://eprint.iacr.org/2006/445.pdf.

[5] L. Budaghyan, C. Carlet, and G. Leander, *Another class of quadratic APN binomials over F_{2^n}: The case n divisible by* 4, in Proceedings of WCC'07, Versailles, France, 2007, pp. 49–58.

[6] L. Budaghyan, C. Carlet, P. Felke, and G. Leander, *An infinite class of quadratic APN functions which are not equivalent to power mappings*, in Proceedings of the ISIT 2006, Seattle, 2006.

[7] L. Budaghyan, C. Carlet, and G. Leander, *Constructing New APN Functions from Known Ones*, preprint, 2007. Available online at http://eprint.iacr.org/2007/063.pdf.

[8] C. Carlet, *Vectorial Boolean functions for cryptography*, to appear in Boolean Methods and Models, P. Hammer and Y. Crama, eds., Cambridge University Press, Cambridge, UK. Available online at http://www-rocq.inria.fr/secret/Claude.Carlet/chap-vectorial-fcts.pdf.

[9] A. Canteaut, P. Charpin, and H. Dobbertin, *Weight divisibility of cyclic codes, highly nonlinear functions on F_{2^m}, and crosscorrelation of maximum-length sequences*, SIAM J. Discrete Math., 13 (2000), pp. 105–138.

[10] C. Carlet, P. Charpin, and V. Zinoviev, *Codes, bent functions and permutations suitable for DES-like cryptosystems*, Des. Codes Cryptogr., 15 (1998), pp. 125–156.

[11] F. Chabaud and S. Vaudenay, *Links between differential and linear cryptanalysis*, Advances in Cryptology (EUROCRYPT'94), Lecture Notes in Comput. Sci. 950, Springer-Verlag, Berlin, 1995, pp. 356–365.

[12] J. Dillon, *Slides from Talk Given at Polynomials over Finite Fields and Applications*, at Banff International Research Station, Banff, AB, Canada, 2006.

[13] Y. Edel, G. Kyureghyan, and A. Pott, *A new APN function which is not equivalent to a power mapping*, IEEE Trans. Inform. Theory, 52 (2006), pp. 744–747.
[14] K. Nyberg, *Differentially uniform mappings for cryptography*, in Advances in Cryptology (EUROCRYPT'93), Lecture Notes in Comput. Sci. 765, Springer, Berlin, 1994, pp. 55–64.
[15] H. M. Trachtenberg, *On the Cross-Correlation Functions of Maximal Linear Sequences*, Ph.D. dissertation, University of Southern California, Los Angeles, 1970.

SIAM J. DISCRETE MATH.
Vol. 23, No. 2, pp. 609–633

AN ARBITRARY STARTING HOMOTOPY-LIKE SIMPLICIAL ALGORITHM FOR COMPUTING AN INTEGER POINT IN A CLASS OF POLYTOPES*

CHUANGYIN DANG†

Abstract. An arbitrary starting homotopy-like simplicial algorithm is developed for computing an integer point in a polytope given by $P = \{x \mid Ax \leq b\}$ satisfying that each row of A has at most one positive entry. The algorithm is derived from an introduction of an integer labeling rule and an application of a triangulation of the space $R^n \times [0,1]$. It consists of two phases, one of which forms an $(n+1)$-dimensional pivoting procedure and the other an n-dimensional pivoting procedure. Starting from an arbitrary integer point in $R^n \times \{0\}$, the algorithm interchanges from one phase to the other, if necessary, and follows a finite simplicial path that either leads to an integer point in the polytope or proves that no such point exists.

Key words. integer point, polytope, integer programming, integer labeling, triangulation, homotopy-like simplicial algorithm, pivoting procedure

AMS subject classification. 90C10

DOI. 10.1137/07069715X

1. Introduction.

The problem we consider is as follows. Determine whether there is an integer point in a polytope given by

$$P = \{x \in R^n \mid Ax \leq b\}, \tag{1}$$

where

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

satisfies that each row of A has at most one positive entry and $b = (b_1, b_2, \ldots, b_m)^\top$. The simultaneous Diophantine approximation problems in [19] and the problem of finding an integer point satisfying a system of monotone inequalities in [13] are special cases of (1). Under some further restrictions, the problem was studied in [27]. The problem is NP-complete (see, e.g., [22, 28]) and very general although it looks special. It is well known that integer programming is equivalent to determining whether there is an integer point in a polytope. By aggregations, integer programming can be reduced in polynomial time to determining whether there is an integer point in a simplex [14]. A simplex is a special polytope given by $P = \{x \mid Ax \leq b\}$ with A being an $(n+1) \times n$ matrix. For any given integer $(n+1) \times n$ matrix A satisfying that $\rho^\top A = 0$ for some positive vector ρ and that any $n \times n$ submatrix of A is nonsingular, a procedure in [23] shows that, by applying the following three elementary column operations to A:

*Received by the editors July 13, 2007; accepted for publication (in revised form) October 28, 2008; published electronically February 6, 2009. This work was partially supported by CERG: CityU 101003 of the Government of Hong Kong SAR.
http://www.siam.org/journals/sidma/23-2/69715.html

†Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong (mecdang@cityu.edu.hk).

1. interchange two columns,
2. multiply a column by -1, and
3. add any integer multiple of a column to another column,

one can transform A in polynomial time into a matrix such that each row has at most one positive entry.

To compute an integer point in a simplex, simplicial algorithms were developed in [4, 5, 6]. The idea of developing a simplicial algorithm for integer programming was stimulated from the seminal work of Scarf [26]. Such an approach to integer programming has its foundations in simplicial algorithms for computing fixed points of a continuous or upper semicontinuous mapping that were originated by Scarf [24] and substantially developed in the literature (see, e.g., [1, 2, 3, 7, 8, 9, 10, 12, 15, 16, 17, 18, 20, 21, 25, 29, 30, 31]). The basic idea of a simplicial algorithm for computing an integer point in a simplex is as follows. The algorithm assigns an integer label to each point of the space and applies a triangulation to subdivide the space into simplices in such a way that every integer point of the space is a vertex of a simplex of the triangulation and every vertex of a simplex of the triangulation is an integer point of the space. Starting from an arbitrary integer point, the algorithm follows a finite simplicial path that either leads to an integer point in the simplex or proves that no such point exists.

In this paper, an arbitrary starting homotopy-like simplicial algorithm is developed for computing an integer point in the polytope given by (1). The algorithm is derived from an introduction of an integer labeling rule and an application of a triangulation of the space $R^n \times [0,1]$. It consists of two phases, one of which forms an $(n+1)$-dimensional pivoting procedure and the other an n-dimensional pivoting procedure. Starting from an arbitrary integer point in $R^n \times \{0\}$, the algorithm interchanges from one phase to the other, if necessary, and follows a finite simplicial path that either leads to an integer point in the polytope or proves that no such point exists.

The rest of this paper is organized as follows. In section 2, we introduce an integer labeling rule and study its properties. In section 3, based on the integer labeling rule and a triangulation of $R^n \times [0,1]$, we develop an algorithm for computing an integer point in the polytope given by (1) and prove its finite convergence.

2. An integer labeling rule and its properties. For $i = 1, 2, \ldots, m$, let $a_i^\top$ denote the ith row of A. Then, $A = (a_1, a_2, \ldots, a_m)^\top$. Let $M = \{1, 2, \ldots, m\}$, $N = \{1, 2, \ldots, n\}$, and $N_0 = \{1, 2, \ldots, n+1\}$. Let $\eta = (\eta_1, \eta_2, \ldots, \eta_n)^\top$ be an arbitrary integer point of R^n, which will be the starting point of the algorithm. We assume that P is bounded and has an interior point. As a result of this assumption, using the well-known separation theorem, one can easily obtain the following lemma.

LEMMA 1. *There is a positive vector* $\rho = (\rho_1, \rho_2, \ldots, \rho_m)^\top$ *satisfying that* $\rho^\top A = 0$.

To implement the algorithm, we need a triangulation of $R^n \times [0,1]$, which subdivides $R^n \times [0,1]$ into simplices in such a way that every integer point of $R^n \times [0,1]$ is a vertex of a simplex of the triangulation, and every vertex of a simplex of the triangulation is an integer point of $R^n \times [0,1]$. There exist several triangulations of $R^n \times [0,1]$ suitable for the purpose in the literature, which include the K_1-triangulation in Freudenthal [11], the J_1-triangulation in Todd [29], and the D_1-triangulation in Dang [2]. The choice of a triangulation of $R^n \times [0,1]$ plays no dominant rule at all in this paper, although the efficiency of the algorithm depends critically on the underlying triangulation. For simplicity of the algorithm, we choose the D_1-triangulation as

an underlying triangulation of the algorithm. For completeness of the algorithm, we also introduce the D_1-triangulation here.

For $j = 1, 2, \ldots, n+1$, let u^j denote the jth unit vector of R^{n+1}. A simplex of the D_1-triangulation of $R^n \times [0,1]$ is the convex hull of $n+2$ vectors, y^0, y^1, $\ldots$, y^{n+1}, given as follows. If $p = 0$, then $y^0 = y$ and $y^k = y + s_{\pi(k)} u^{\pi(k)}$, $k = 1, 2, \ldots, n+1$, and, if $p \geq 1$, then $y^0 = y + s$, $y^k = y^{k-1} - s_{\pi(k)} u^{\pi(k)}$, $k = 1, 2, \ldots, p-1$, and $y^k = y + s_{\pi(k)} u^{\pi(k)}$, $k = p, p+1, \ldots, n+1$, where $y = (\eta, 0)^\top + (z_1, z_2, \ldots, z_n, 0)^\top$ with z_i being an even number for $i = 1, 2, \ldots, n$, $\pi = (\pi(1), \pi(2), \ldots, \pi(n+1))$ a permutation of the elements of $N_0 = \{1, 2, \ldots, n+1\}$, $s = (s_1, s_2, \ldots, s_{n+1})^\top$ a sign vector with every component being a number in $\{-1, 1\}$ and $s_{n+1} = 1$, and p an integer with $0 \leq p \leq n$. Let D_1 be the set of all such simplices. Then, D_1 is a triangulation of $R^n \times [0,1]$. Since a simplex of the D_1-triangulation is determined by y, π, s, and p, we use $D_1(y, \pi, s, p)$ to denote it.

We say that two simplices of D_1 are adjacent if they share a common facet. We show in the following how to generate all the adjacent simplices of a simplex of the D_1-triangulation of $R^n \times [0,1]$. For a given simplex $\sigma = D_1(y, \pi, s, p)$ with vertices y^0, y^1, $\ldots$, y^{n+1}, its adjacent simplex opposite to a vertex, say y^i, is given by $D_1(\bar{y}, \bar{\pi}, \bar{s}, \bar{p})$, where $\bar{y}$, $\bar{\pi}$, $\bar{s}$, and $\bar{p}$ are generated according to the pivot rules given in Table 1.

TABLE 1
Pivot rules of the D_1-triangulation of $R^n \times [0,1]$.

i				$\bar{y}$	$\bar{s}$	$\bar{\pi}$	$\bar{p}$
0	$p \leq 1$			y	s	π	$1-p$
	$2 \leq p$		$\pi(1) \neq n+1$	y	$s - 2s_{\pi(1)} u^{\pi(1)}$	π	p
			$\pi(1) = n+1$	The facet opposite to y^i is contained in $R^n \times \{0\}$			
$1 \leq i$		$p = 0$	$\pi(i) \neq n+1$	y	$s - 2s_{\pi(i)} u^{\pi(i)}$	π	p
			$\pi(i) = n+1$	The facet opposite to y^i is contained in $R^n \times \{0\}$			
		$i < p-1$		y	s	π^1	p
		$i = p-1$		y	s	π	$p-1$
	$p-1 < i$	$1 \leq p < n$		y	s	π^2	$p+1$
	$i = n$	$1 \leq p = n$	$\pi(n+1) \neq n+1$	$y + 2s_{\pi(n+1)} u^{\pi(n+1)}$	$s - 2s_{\pi(n+1)} u^{\pi(n+1)}$	π	p
			$\pi(n+1) = n+1$	The facet opposite to y^i is contained in $R^n \times \{1\}$			
	$i = n+1$	$1 \leq p = n$	$\pi(n) \neq n+1$	$y + 2s_{\pi(n)} u^{\pi(n)}$	$s - 2s_{\pi(n)} u^{\pi(n)}$	π	p
			$\pi(n) = n+1$	The facet opposite to y^i is contained in $R^n \times \{1\}$			

$\pi^1 = (\pi(1), \ldots, \pi(i+1), \pi(i), \ldots, \pi(n+1))$,
$\pi^2 = (\pi(1), \ldots, \pi(p-1), \pi(i), \pi(p), \ldots, \pi(i-1), \pi(i+1), \ldots, \pi(n+1))$.

Let $\mathcal{D}_1$ be the set of faces of simplices of D_1. A q-dimensional simplex of $\mathcal{D}_1$ with vertices y^0, $y^1, \ldots, y^q$ is denoted by $\langle y^0, y^1, \ldots, y^q \rangle$. The restriction of the D_1-triangulation of $R^n \times [0,1]$ on $R^n \times \{0\}$ is given by

$$D_1 | R^n \times \{0\} = \{\sigma \in \mathcal{D}_1 \mid \sigma \subset R^n \times \{0\} \text{ and } \dim(\sigma) = n\},$$

and the restriction of the D_1-triangulation of $R^n \times [0,1]$ on $R^n \times \{1\}$ is given by

$$D_1 | R^n \times \{1\} = \{\sigma \in \mathcal{D}_1 \mid \sigma \subset R^n \times \{1\} \text{ and } \dim(\sigma) = n\},$$

where $\dim(\cdot)$ stands for the dimension of a set.

From the definition of the D_1-triangulation of $R^n \times [0,1]$, we know that $D_1 | R^n \times \{0\}$ is the same as the D_1-triangulation of R^n. For $j = 1, 2, \ldots, n$, let u^j denote the jth unit vector of R^n. A simplex of the D_1-triangulation of R^n is the convex hull of $n+1$

vectors, $y^0, y^1, \ldots, y^n$, given as follows. If $p = 0$, then $y^0 = y$ and $y^k = y + s_{\pi(k)}u^{\pi(k)}$, $k = 1, 2, \ldots, n$, and, if $p \geq 1$, then $y^0 = y + s$, $y^k = y^{k-1} - s_{\pi(k)}u^{\pi(k)}$, $k = 1, 2, \ldots, p-1$, and $y^k = y + s_{\pi(k)}u^{\pi(k)}$, $k = p, p+1, \ldots, n$, where $y = \eta + z$ with every component of z being an even number, $\pi = (\pi(1), \pi(2), \ldots, \pi(n))$ a permutation of the elements of $N = \{1, 2, \ldots, n\}$, s a sign vector with every component being a number in $\{-1, 1\}$, and p an integer with $0 \leq p \leq n-1$. Let D_1 be the set of all such simplices. Then, D_1 is a triangulation of R^n. Since a simplex of the D_1-triangulation is determined by y, π, s, and p, we use $D_1(y, \pi, s, p)$ to denote it. How to generate all the adjacent simplices of a simplex of the D_1-triangulation of R^n is as follows. For a given simplex $\sigma = D_1(y, \pi, s, p)$ with vertices y^0, $y^1, \ldots, y^n$, its adjacent simplex opposite to a vertex, say y^i, is given by $D_1(\bar{y}, \bar{\pi}, \bar{s}, \bar{p})$, where $\bar{y}$, $\bar{\pi}$, $\bar{s}$, and $\bar{p}$ are generated according to the pivot rules given in Table 2.

TABLE 2
Pivot rules of the D_1-triangulation of R^n.

i			$\bar{y}$	$\bar{s}$	$\bar{\pi}$	$\bar{p}$
0	$n = 1$		$y + 2s_{\pi(1)}u^{\pi(1)}$	$s - 2s_{\pi(1)}u^{\pi(1)}$	π	p
	$n \geq 2$	$p \leq 1$	y	s	π	$1 - p$
		$2 \leq p$	y	$s - 2s_{\pi(1)}u^{\pi(1)}$	π	p
$1 \leq i$		$p = 0$	y	$s - 2s_{\pi(i)}u^{\pi(i)}$	π	p
		$i < p - 1$	y	s	π^1	p
		$i = p - 1$	y	s	π	$p - 1$
	$p - 1 < i$	$1 \leq p < n - 1$	y	s	π^2	$p + 1$
	$i = n - 1$	$1 \leq p = n - 1$	$y + 2s_{\pi(n)}u^{\pi(n)}$	$s - 2s_{\pi(n)}u^{\pi(n)}$	π	p
	$i = n$	$1 \leq p = n - 1$	$y + 2s_{\pi(n-1)}u^{\pi(n-1)}$	$s - 2s_{\pi(n-1)}u^{\pi(n-1)}$	π	p

$\pi^1 = (\pi(1), \ldots, \pi(i+1), \pi(i), \ldots, \pi(n))$,
$\pi^2 = (\pi(1), \ldots, \pi(p-1), \pi(i), \pi(p), \ldots, \pi(i-1), \pi(i+1), \ldots, \pi(n))$.

From the definition of the D_1-triangulation of $R^n \times [0, 1]$, we know that $D_1 | R^n \times \{1\}$ is the same as the J_1-triangulation of R^n. For $j = 1, 2, \ldots, n$, let u^j denote the jth unit vector of R^n. A simplex of the J_1-triangulation of R^n is the convex hull of $n + 1$ vectors, y^0, $y^1, \ldots, y^n$, given as follows. $y^0 = y + s$, $y^k = y^{k-1} - s_{\pi(k)}u^{\pi(k)}$, $k = 1, 2, \ldots, n$, where $y = \eta + z$ with every component of z being an even number, $\pi = (\pi(1), \pi(2), \ldots, \pi(n))$ a permutation of the elements of $N = \{1, 2, \ldots, n\}$, and s a sign vector with every component being a number in $\{-1, 1\}$. Let J_1 be the set of all such simplices. Then, J_1 is a triangulation of R^n. Since a simplex of the J_1-triangulation is determined by y, π, and s, we use $J_1(y, \pi, s)$ to denote it. How to generate all the adjacent simplices of a simplex of the J_1-triangulation of R^n is as follows. For a given simplex $\sigma = J_1(y, \pi, s)$ with vertices y^0, $y^1, \ldots, y^n$, its adjacent simplex opposite to a vertex, say y^i, is given by $J_1(\bar{y}, \bar{\pi}, \bar{s})$, where $\bar{y}$, $\bar{\pi}$, and $\bar{s}$ are generated according to the pivot rules given in Table 3.

TABLE 3
Pivot rules of the J_1-triangulation of R^n.

i	$\bar{y}$	$\bar{s}$	$\bar{\pi}$
0	y	$s - 2s_{\pi(1)}u^{\pi(1)}$	π
$0 < i < n$	y	s	$(\pi(1), \ldots, \pi(i+1), \pi(i), \ldots, \pi(n))$
n	$y + 2s_{\pi(n)}u^{\pi(n)}$	$s - 2s_{\pi(n)}u^{\pi(n)}$	π

For $\sigma \in \mathcal{D}_1$, let $\text{grid}(\sigma) = \max\{\|x - y\| \mid x \in \sigma \text{ and } y \in \sigma\}$, where $\|\cdot\|$ denotes the infinity norm. We define $\text{mesh}(D_1) = \max_{\sigma \in \mathcal{D}_1} \text{grid}(\sigma)$. Then, $\text{mesh}(D_1) = 1$.

For any $x \in R^n$, we define

$$f(x) = (f_1(x), f_2(x), \ldots, f_n(x))^\top = \begin{cases} 0 \in R^n & \text{if } x \in P, \\ \sum_{j \in J(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j & \text{otherwise,} \end{cases}$$

where $J(x) = \{j \in M \mid a_j^\top x - b_j > 0\}$.

Let

$$d = A\eta.$$

For any $x \in R^n$, we define

$$f^0(x) = (f_1^0(x), f_2^0(x), \ldots, f_n^0(x))^\top = \begin{cases} 0 \in R^n & \text{if } Ax \leq d, \\ \sum_{j \in J^0(x)} \frac{a_j^\top x - d_j}{a_j^\top a_j} a_j & \text{otherwise,} \end{cases}$$

where $J^0(x) = \{j \in M \mid a_j^\top x - d_j > 0\}$.

LEMMA 2. *The integer point η is a unique point satisfying that $Ax \leq d$.*

Proof. From the boundedness assumption on P, we know that A has a nonsingular $n \times n$ submatrix. Then, $x = \eta$ for any $x \in R^n$ with $Ax = A\eta$. Suppose that there is a point $y \in R^n$ such that $Ay \leq d$ and $y \neq \eta$. Then, there is at least one index $i \in M$ satisfying that $a_i^\top y < d_i$. By Lemma 1, we derive that

$$0 = \rho^\top A(y - \eta) < 0.$$

A contradiction occurs. The lemma follows. □

From the definitions of $f(x)$ and $f^0(x)$, we obtain that

$$f(x) = \left(\sum_{j \in J(x)} \frac{a_j a_j^\top}{a_j^\top a_j} \right) x - \sum_{j \in J(x)} \frac{b_j}{a_j^\top a_j} a_j$$

and

$$f^0(x) = \left(\sum_{j \in J^0(x)} \frac{a_j a_j^\top}{a_j^\top a_j} \right) x - \sum_{j \in J^0(x)} \frac{d_j}{a_j^\top a_j} a_j.$$

Clearly, both $f : R^n \to R^n$ and $f^0 : R^n \to R^n$ are continuous piecewise linear mappings, each of which is composed of a finite number of linear pieces since there is only a finite number of different $J(x)$'s and $J^0(x)$'s on R^n.

LEMMA 3. *For any x and y in R^n,*

$$\|f(x) - f(y)\| \leq m\|x - y\| \tag{2}$$

and

$$\|f^0(x) - f^0(y)\| \leq m\|x - y\|. \tag{3}$$

Proof. Let x and y be any two points in R^n. Then, for any $j \in J(x)$ and $j \notin J(y)$,

$$a_j^\top (x - y) \geq a_j^\top x - b_j > 0,$$

and for any $j \notin J(x)$ and $j \in J(y)$,

$$a_j^\top (y - x) \geq a_j^\top y - b_j > 0.$$

Thus,

$$\begin{aligned}
&\|f(x) - f(y)\| = \left\| \textstyle\sum_{j\in J(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j - \sum_{j\in J(y)} \frac{a_j^\top y - b_j}{a_j^\top a_j} a_j \right\| \\
&= \left\| \textstyle\sum_{j\in J(x)\cap J(y)} \frac{a_j^\top x - a_j^\top y}{a_j^\top a_j} a_j + \sum_{j\notin J(y)}^{j\in J(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j - \sum_{j\notin J(x)}^{j\in J(y)} \frac{a_j^\top y - b_j}{a_j^\top a_j} a_j \right\| \\
&\leq \left\| \textstyle\sum_{j\in J(x)\cap J(y)} \frac{a_j^\top (x-y)}{a_j^\top a_j} a_j \right\| + \sum_{j\notin J(y)}^{j\in J(x)} \frac{a_j^\top (x-y)}{a_j^\top a_j} \|a_j\| + \sum_{j\notin J(x)}^{j\in J(y)} \frac{a_j^\top (y-x)}{a_j^\top a_j} \|a_j\| \\
&\leq \textstyle\sum_{j\in J(x)\cap J(y)} \frac{\|a_j\|^2 \|x-y\|}{a_j^\top a_j} + \sum_{j\notin J(y)}^{j\in J(x)} \frac{\|a_j\|^2 \|x-y\|}{a_j^\top a_j} + \sum_{j\notin J(x)}^{j\in J(y)} \frac{\|a_j\|^2 \|x-y\|}{a_j^\top a_j} \\
&= (|J(x) \cap J(y)| + |J(x) \backslash J(y)| + |J(y) \backslash J(x)|) \|x - y\| \\
&\leq m \|x - y\|,
\end{aligned}$$

where $|\cdot|$ denotes the cardinality of a set. Similarly, one can derive that $\|f^0(x) - f^0(y)\| \leq m\|x - y\|$ for any x and y in R^n. The lemma follows. ∎

Clearly, for any $y \in R^n$, we have, for any $j \in J(y)$ and $j \notin J^0(y)$,

$$0 < a_j^\top y - b_j \leq d_j - b_j,$$

and for any $j \notin J(y)$ and $j \in J^0(y)$,

$$0 < a_j^\top y - d_j \leq b_j - d_j.$$

Thus, for any $y \in R^n$, we have

$$\begin{aligned}
&\|f^0(y) - f(y)\| \\
&= \left\| \textstyle\sum_{j\in J^0(y)\cap J(y)} \frac{b_j - d_j}{a_j^\top a_j} a_j + \sum_{j\notin J(y)}^{j\in J^0(y)} \frac{a_j^\top y - d_j}{a_j^\top a_j} a_j - \sum_{j\notin J^0(y)}^{j\in J(y)} \frac{a_j^\top y - b_j}{a_j^\top a_j} a_j \right\| \\
&\leq \textstyle\sum_{j\in J^0(y)\cap J(y)} \frac{|b_j - d_j|}{a_j^\top a_j} \|a_j\| + \sum_{j\notin J(y)}^{j\in J^0(y)} \frac{a_j^\top y - d_j}{a_j^\top a_j} \|a_j\| + \sum_{j\notin J^0(y)}^{j\in J(y)} \frac{a_j^\top y - b_j}{a_j^\top a_j} \|a_j\| \qquad (4) \\
&\leq \textstyle\sum_{j\in J^0(y)\cap J(y)} \frac{|b_j - d_j|}{\|a_j\|} + \sum_{j\notin J(y)}^{j\in J^0(y)} \frac{b_j - d_j}{\|a_j\|} + \sum_{j\notin J^0(y)}^{j\in J(y)} \frac{d_j - b_j}{\|a_j\|} \\
&\leq \textstyle\sum_{j\in M} \frac{|d_j - b_j|}{\|a_j\|}.
\end{aligned}$$

Therefore, for any x and y in R^n, as a result of (2) and (4), we have

$$\begin{aligned}
\|f(x) - f^0(y)\| &= \|f(x) - f(y) + f(y) - f^0(y)\| \\
&\leq \|f(x) - f(y)\| + \|f(y) - f^0(y)\| \qquad (5) \\
&\leq m\|x - y\| + \textstyle\sum_{j\in M} \frac{|d_j - b_j|}{\|a_j\|}.
\end{aligned}$$

For any $(x, t) \in R^n \times [0, 1]$, we define

$$h(x, t) = (h_1(x, t), h_2(x, t), \ldots, h_n(x, t))^\top = tf(x) + (1 - t)f^0(x).$$

Then, for any (x, t_1) and (y, t_2) in $R^n \times [0, 1]$, as a result of (2), (3), and (5), we have

$$
\begin{aligned}
&\|h(x,t_1) - h(y,t_2)\| \\
&= \|t_1 f(x) + (1-t_1) f^0(x) - (t_2 f(y) + (1-t_2) f^0(y))\| \\
&= \|t_2(f(x) - f(y)) + (1-t_1)(f^0(x) - f^0(y)) + (t_1 - t_2)(f(x) - f^0(y))\| \\
&\le \|f(x) - f(y)\| + \|f^0(x) - f^0(y)\| + \|f(x) - f^0(y)\| \\
&\le m\|x-y\| + m\|x-y\| + m\|x-y\| + \textstyle\sum_{j\in M} \frac{|d_j - b_j|}{\|a_j\|} \\
&= 3m\|x-y\| + \textstyle\sum_{j\in M} \frac{|d_j - b_j|}{\|a_j\|}.
\end{aligned} \tag{6}
$$

LEMMA 4. *For any given $x^* \in R^n$, as $\|x\| \to \infty$, $\frac{(x-x^*)^\top f(x)}{\|x\|} \to \infty$, $\frac{(x-x^*)^\top f^0(x)}{\|x\|} \to \infty$ and $\frac{(x-x^*)^\top h(x,t)}{\|x\|} \to \infty$ for any $t \in [0,1]$.*

Proof. Let x^0 be any given point of P. Because P is bounded, there is a ball $B(x^0, r)$ strictly containing P. Let $S(x^0, r)$ be the sphere of the ball. Then, for any $x \notin B(x^0, r)$, there are some point $y \in S(x^0, r)$ and some number $\rho > 1$ satisfying that $x = x^0 + \rho(y - x^0)$. Thus, for any k,

$$
\begin{aligned}
a_k^\top x - b_k &= a_k^\top (x^0 + \rho(y - x^0)) - b_k \\
&= \rho(a_k^\top y - b_k) + (\rho - 1)(b_k - a_k^\top x^0) \\
&\ge \rho(a_k^\top y - b_k),
\end{aligned}
$$

where the last inequality comes from $b_k \ge a_k^\top x^0$ and $\rho > 1$. Therefore, for any $k \in J(y)$, as a result of $a_k^\top y - b_k > 0$, we obtain that $a_k^\top x - b_k$ approaches infinity as $\rho \to \infty$. Observe that, for any $y \in S(x^0, r)$, $J(y)$ is not empty and that, for any $x \notin P$ with $x \ne 0$,

$$
\begin{aligned}
\tfrac{(x-x^*)^\top f(x)}{\|x\|} &= \textstyle\sum_{j\in J(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j \|x\|} a_j^\top (x - x^*) \\
&= \textstyle\sum_{j\in J(x)} \frac{(a_j^\top x - b_j)^2}{a_j^\top a_j \|x\|} + \sum_{j\in J(x)} \frac{(a_j^\top x - b_j)(b_j - a_j^\top x^*)}{a_j^\top a_j \|x\|}.
\end{aligned}
$$

Thus, $\frac{(x-x^*)^\top f(x)}{\|x\|} \to \infty$ as $\|x\| \to \infty$. Similarly, one can show that $\frac{(x-x^*)^\top f^0(x)}{\|x\|} \to \infty$ as $\|x\| \to \infty$.

For any $t \in [0, 1]$, since

$$
\frac{(x-x^*)^\top h(x,t)}{\|x\|} = t\frac{(x-x^*)^\top f(x)}{\|x\|} + (1-t)\frac{(x-x^*)^\top f^0(x)}{\|x\|},
$$

hence, $\frac{(x-x^*)^\top h(x,t)}{\|x\|} \to \infty$ as $\|x\| \to \infty$. The lemma follows. ∎

From an application of $f(x)$ and $f^0(x)$, we obtain the following integer labeling rule for assigning an integer to each integer point of $R^n \times [0, 1]$.

DEFINITION 1. *For $(x, 1) \in R^n \times \{1\}$, we assign to $(x, 1)$ an integer $l(x, 1)$ given by $l(x, 1) = 0$ if $f(x) = 0$, and*

$$
l(x,1) = \begin{cases} \max\{k \mid f_k(x) = \max_{1\le j\le n} f_j(x)\} & \text{if } f_j(x) > 0 \text{ for some } j \in N, \\ n+1 & \text{if } f(x) \le 0 \text{ and } f(x) \ne 0. \end{cases}
$$

For $(x,0) \in R^n \times \{0\}$, we assign to $(x,0)$ an integer $l(x,0)$ given by

$$l(x,0) = \begin{cases} \max\{k \mid f_k^0(x) = \max_{1\le j\le n} f_j^0(x)\} & \text{if } f_j^0(x) > 0 \text{ for some } j \in N, \\ n+1 & \text{if } f^0(x) \le 0. \end{cases}$$

From Definition 1, one can see that there is no point in the artificial level $R^n \times \{0\}$ that carries integer label 0.

The next definition gives us a few notations that will be used in our further developments.

DEFINITION 2.

- *A q-dimensional simplex $\sigma = \langle y^0, y^1, \ldots, y^q\rangle$ of $\mathcal{D}_1$ is complete if $l(y^i) \neq l(y^j)$ for $0 \le i < j \le q$, and $l(y^k) \neq 0$, $k = 0, 1, \ldots, q$.*
- *A q-dimensional simplex $\sigma = \langle y^0, y^1, \ldots, y^q\rangle$ of $\mathcal{D}_1$ is 0-complete if $l(y^i) \neq l(y^j)$ for $0 \le i < j \le q$, and there is some k satisfying that $l(y^k) = 0$.*
- *A q-dimensional simplex $\sigma = \langle y^0, y^1, \ldots, y^q\rangle$ of $\mathcal{D}_1$ is almost complete if labels of $q+1$ vertices of σ consist of q different nonzero integers.*

From Definition 2, it is easy to see the following lemma.

LEMMA 5. *An almost complete simplex has exactly two complete facets.*

For any $y = (y_1, y_2, \ldots, y_{n+1})^\top \in R^n \times [0,1]$, we define

$$p(y) = (y_1, y_2, \ldots, y_n)^\top.$$

THEOREM 1. *There is a finite number of complete n-dimensional simplices in $\mathcal{D}_1$.*

Proof. Let $\sigma = \langle y^0, y^1, \ldots, y^n\rangle$ be a complete n-dimensional simplex in $\mathcal{D}_1$. Without loss of generality, we assume that $l(y^0) = n+1$ and $l(y^i) = i$, $i = 1, 2, \ldots, n$. Then, from Definition 1, we know that

$$h(y^0) \le 0 \tag{7}$$

and

$$h_i(y^i) > 0,\ i = 1, 2, \ldots, n. \tag{8}$$

Let y be an arbitrary point of σ. Then, as a result of (7), (8), (6), and $\mathrm{mesh}(D_1) = 1$, we have, for $i = 1, 2, \ldots, n$,

$$\begin{aligned} h_i(y) &= h_i(y) - h_i(y^0) + h_i(y^0) \\ &\le h_i(y) - h_i(y^0) \\ &\le 3m\|y - y^0\| + \textstyle\sum_{j\in M} \frac{|d_j - b_j|}{\|a_j\|} \\ &\le 3m + \textstyle\sum_{j\in M} \frac{|d_j - b_j|}{\|a_j\|} \end{aligned} \tag{9}$$

and

$$\begin{aligned} h_i(y) &= h_i(y) - h_i(y^i) + h_i(y^i) \\ &\ge h_i(y) - h_i(y^i) \\ &\ge -\Big(3m\|y - y^i\| + \textstyle\sum_{j\in M} \frac{|d_j - b_j|}{\|a_j\|}\Big) \\ &\ge -\Big(3m + \textstyle\sum_{j\in M} \frac{|d_j - b_j|}{\|a_j\|}\Big). \end{aligned} \tag{10}$$

Thus, (9) and (10) together implies that σ is contained in

$$W_c = \left\{ y \in R^n \times [0,1] \mid \|h(y)\| \le 3m + \sum_{j \in M} \frac{|d_j - b_j|}{\|a_j\|} \right\}.$$

Clearly, for any $y \in W_c$,

$$-|y_i|\left(3m + \sum_{j \in M} \frac{|d_j - b_j|}{\|a_j\|}\right) \le y_i h_i(y) \le |y_i|\left(3m + \sum_{j \in M} \frac{|d_j - b_j|}{\|a_j\|}\right),$$

$i = 1, 2, \ldots, n$. Thus, for any $y \in W_c$,

$$-\left(3m + \sum_{j \in M} \frac{|d_j - b_j|}{\|a_j\|}\right) \sum_{i=1}^{n} |y_i| \le p(y)^\top h(y) \le \left(3m + \sum_{j \in M} \frac{|d_j - b_j|}{\|a_j\|}\right) \sum_{i=1}^{n} |y_i|.$$

Therefore, for any $y \in W_c$,

$$-\left(3m + \sum_{j \in M} \frac{|d_j - b_j|}{\|a_j\|}\right) \le \frac{p(y)^\top h(y)}{\|p(y)\|_1} \le 3m + \sum_{j \in M} \frac{|d_j - b_j|}{\|a_j\|}, \tag{11}$$

where $\|p(y)\|_1 = \sum_{i=1}^{n} |y_i|$. Combining Lemma 4 and (11) together, one can see that W_c is bounded. The theorem follows. □

As a corollary of Theorem 1, we have the following.

COROLLARY 1. *There is a finite number of almost complete* $(n+1)$*-dimensional simplices in* D_1.

LEMMA 6. *It holds that*

$$\bar{x} = (\max\{x_1^1, x_1^2\}, \max\{x_2^1, x_2^2\}, \ldots, \max\{x_n^1, x_n^2\})^\top \in P$$

if $x^1 = (x_1^1, x_2^1, \ldots, x_n^1)^\top \in P$ *and* $x^2 = (x_1^2, x_2^2, \ldots, x_n^2)^\top \in P$.

Proof. Let $x^1 = (x_1^1, x_2^1, \ldots, x_n^1)^\top$ and $x^2 = (x_1^2, x_2^2, \ldots, x_n^2)^\top$ be two arbitrary points of P. Let $\bar{x} = (\max\{x_1^1, x_1^2\}, \max\{x_2^1, x_2^2\}, \ldots, \max\{x_n^1, x_n^2\})^\top$. Let j be an arbitrary index of M. From the assumption on A, we know that each row of A has at most one positive entry. If $a_j \le 0$, then

$$a_j^\top \bar{x} \le a_j^\top x^1 \le b_j.$$

Consider a_j with a positive entry, say a_{ji}. Without loss of generality, we assume that $\bar{x}_i = x_i^1$. Since $a_{jk} \le 0$ for any $k \ne i$,

$$a_j^\top \bar{x} = a_{ji}\bar{x}_i + \sum_{k \ne i} a_{jk}\bar{x}_k = a_{ji}x_i^1 + \sum_{k \ne i} a_{jk}\bar{x}_k \le a_{ji}x_i^1 + \sum_{k \ne i} a_{jk}x_k^1 = a_j^\top x^1 \le b_j.$$

Thus, $\bar{x} \in P$. The lemma follows. □

Let $x^{\max}$ denote the unique solution of $\max_{x \in P} e^\top x$.

LEMMA 7. *For any* $x \in P$, $x \le x^{\max}$.

Proof. Suppose that there is a point $\hat{x} \in P$ satisfying $\hat{x}_q > x_q^{\max}$ for some $q \in N$. Let $\bar{x} = (\max\{\hat{x}_1, x_1^{\max}\}, \max\{\hat{x}_2, x_2^{\max}\}, \ldots, \max\{\hat{x}_n, x_n^{\max}\})^\top$. Then, from Lemma 6, we know that $\bar{x} \in P$. Clearly, $e^\top \bar{x} > e^\top x^{\max}$, which contradicts that $e^\top x^{\max} = \max_{x \in P} e^\top x$. The lemma follows. □

LEMMA 8. *Let*

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n1} & q_{n2} & \cdots & q_{nn} \end{pmatrix}$$

be a matrix such that $q_{ij} \leq 0$ *for any* $i \neq j$ *and* $q_{ii} > 0$, $i = 1, 2, \ldots, n$. *If there is some* $\rho = (\rho_1, \rho_2, \ldots, \rho_n)^\top > 0$ *satisfying that* $\rho^\top Q > 0$, *then* Q *is nonsingular and* $Q^{-1} \geq 0$.

Proof. We prove the lemma by the mathematical induction.

1. When $n = 1$, $Q = (q_{11})$. Since $q_{11} > 0$, Q is nonsingular and $Q^{-1} = (1/q_{11}) \geq 0$. The lemma is true.
2. Assume that the lemma is true for $1 \leq k \leq n-1$. Consider $k = n$. Let

$$U = \begin{pmatrix} 1 & & \\ -q_{21}/q_{11} & 1 & \\ \vdots & & \ddots \\ -q_{n1}/q_{11} & & 1 \end{pmatrix}$$

and

$$W = \begin{pmatrix} 1 & -q_{12}/q_{11} & \cdots & -q_{1n}/q_{11} \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}.$$

The inverse matrix of U is given by

$$U^{-1} = \begin{pmatrix} 1 & & \\ q_{21}/q_{11} & 1 & \\ \vdots & & \ddots \\ q_{n1}/q_{11} & & 1 \end{pmatrix}.$$

Note that $-q_{i1}/q_{11} \geq 0$ and $-q_{1i}/q_{11} \geq 0$, $i = 2, 3, \ldots, n$. Multiplying U to the left-hand side of Q, we obtain that

$$UQ = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ 0 & q_{22} - \frac{q_{21}q_{12}}{q_{11}} & \cdots & q_{2n} - \frac{q_{21}q_{1n}}{q_{11}} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & q_{n2} - \frac{q_{n1}q_{12}}{q_{11}} & \cdots & q_{nn} - \frac{q_{n1}q_{1n}}{q_{11}} \end{pmatrix}.$$

Clearly, all the entries of UQ except its diagonal entries are nonpositive. Multiplying ρ to U^{-1}, we obtain that

$$\rho^\top U^{-1} = \left(\rho_1 + \sum_{i=2}^{n} \frac{q_{i1}\rho_i}{q_{11}}, \rho_2, \ldots, \rho_n\right)^\top.$$

From $0 < q_{11}$ and $0 < \rho^\top Q = (\rho^\top U^{-1})(UQ)$, we derive that

$$\rho_1 + \sum_{i=2}^{n} \frac{q_{i1}\rho_i}{q_{11}} > 0,$$

and all the diagonal entries of UQ are positive. Let $\bar{\rho} = (\rho_2, \ldots, \rho_n)^\top$. By deleting the first row and the first column of UQ, we obtain an $(n-1)\times(n-1)$ matrix,

$$\bar{Q} = \begin{pmatrix} q_{22} - \frac{q_{21}q_{12}}{q_{11}} & \cdots & q_{2n} - \frac{q_{21}q_{1n}}{q_{11}} \\ \vdots & \ddots & \vdots \\ q_{n2} - \frac{q_{n1}q_{12}}{q_{11}} & \cdots & q_{nn} - \frac{q_{n1}q_{1n}}{q_{11}} \end{pmatrix}.$$

Since $0 < \bar{\rho}$, $0 < \bar{\rho}^\top \bar{Q}$, and $\bar{Q}$ is an $(n-1) \times (n-1)$ matrix, it follows from the hypothesis that $\bar{Q}$ is nonsingular and $\bar{Q}^{-1} \geq 0$. Multiplying W to the right-hand side of UQ, we obtain

$$UQW = \begin{pmatrix} q_{11} & 0 \\ 0 & \bar{Q} \end{pmatrix}.$$

Thus, Q is nonsingular and

$$Q^{-1} = W \begin{pmatrix} 1/q_{11} & 0 \\ 0 & \bar{Q}^{-1} \end{pmatrix} U.$$

Therefore, $Q^{-1} \geq 0$ since $q_{11} > 0$, $\bar{Q}^{-1} \geq 0$, $U \geq 0$, and $W \geq 0$. The lemma follows. □

As a corollary of Lemma 8, we obtain the following result.

COROLLARY 2. *For any $x \in R^n$, if $0 < f(x)$, then $0 < x - x^0$ for any $x^0 \in P$.*

Proof. Let $J_i(x) = \{j \in J(x) \mid a_{ji} > 0\}$, $i = 1, 2, \ldots, n$, and $J_{n+1}(x) = \{j \in J(x) \mid a_j \leq 0\}$. Then, $J_1(x), J_2(x), \ldots, J_{n+1}(x)$ forms a partition of $J(x)$. Since $f(x) > 0$, hence, $J_i(x) \neq \emptyset$, $i = 1, 2, \ldots, n$. Let

$$r_i(x) = \sum_{j \in J_i(x)} \frac{(a_j^\top x - b_j)^2}{a_j^\top a_j},$$

$i = 1, 2, \ldots, n$, and $r(x) = (r_1(x), r_2(x), \ldots, r_n(x))^\top$. Clearly, $r(x) > 0$. Let

$$\bar{a}_i(x) = \frac{\sum_{j \in J_i(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j}{r_i(x)},$$

$i = 1, 2, \ldots, n$, and $\bar{A}(x) = (\bar{a}_1(x), \bar{a}_2(x), \ldots, \bar{a}_n(x))$. Since $0 < f(x)$ and $\sum_{j \in J_{n+1}(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j \leq 0$, hence,

$$\begin{aligned} 0 &< f(x) - \textstyle\sum_{j \in J_{n+1}(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j = \sum_{j \in J(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j - \sum_{j \in J_{n+1}(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j \\ &= \textstyle\sum_{i=1}^n \sum_{j \in J_i(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j = \sum_{i=1}^n \frac{\sum_{j \in J_i(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j}{r_i(x)} r_i(x) \\ &= \textstyle\sum_{i=1}^n \bar{a}_i(x) r_i(x) = \bar{A}(x) r(x). \end{aligned}$$

From the definition of $\bar{A}(x)$, we know that each row and each column of $\bar{A}(x)$ have exactly one positive entry. Then, by Lemma 8, we obtain that $\bar{A}(x)$ is nonsingular and $\bar{A}(x)^{-1} \geq 0$. From the definition of $r(x)$, we obtain that

$$r(x) = \begin{pmatrix} \sum_{j\in J_1(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j}(a_j^\top (x - x^0) + a_j^\top x^0 - b_j) \\ \sum_{j\in J_2(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j}(a_j^\top (x - x^0) + a_j^\top x^0 - b_j) \\ \vdots \\ \sum_{j\in J_n(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j}(a_j^\top (x - x^0) + a_j^\top x^0 - b_j) \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{j\in J_1(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top \\ \sum_{j\in J_2(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top \\ \vdots \\ \sum_{j\in J_n(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top \end{pmatrix} (x - x^0) + \begin{pmatrix} \sum_{j\in J_1(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j}(a_j^\top x^0 - b_j) \\ \sum_{j\in J_2(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j}(a_j^\top x^0 - b_j) \\ \vdots \\ \sum_{j\in J_n(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j}(a_j^\top x^0 - b_j) \end{pmatrix}.$$

Let

$$s(x^0) = \begin{pmatrix} \sum_{j\in J_1(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j}(a_j^\top x^0 - b_j) \\ \sum_{j\in J_2(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j}(a_j^\top x^0 - b_j) \\ \vdots \\ \sum_{j\in J_n(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j}(a_j^\top x^0 - b_j) \end{pmatrix}.$$

Then, $s(x^0) \leq 0$ since $x^0 \in P$. Thus,

$$0 < r(x) - s(x^0) = \begin{pmatrix} \sum_{j\in J_1(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top \\ \sum_{j\in J_2(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top \\ \vdots \\ \sum_{j\in J_n(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top \end{pmatrix} (x - x^0).$$

Let

$$R(x) = \begin{pmatrix} r_1(x) & & & \\ & r_2(x) & & \\ & & \ddots & \\ & & & r_n(x) \end{pmatrix}.$$

Then,

$$R(x)\bar{A}(x)^\top = \begin{pmatrix} \sum_{j\in J_1(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top \\ \sum_{j\in J_2(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top \\ \vdots \\ \sum_{j\in J_n(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top \end{pmatrix}.$$

Therefore, since $\bar{A}(x)^{-1} \geq 0$,

$$\begin{aligned} x - x^0 &= \begin{pmatrix} \sum_{j\in J_1(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top \\ \sum_{j\in J_2(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top \\ \vdots \\ \sum_{j\in J_n(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top \end{pmatrix}^{-1} (r(x) - s(x^0)) \\ &= (\bar{A}(x)^{-1})^\top R(x)^{-1}(r(x) - s(x^0)) > 0. \end{aligned}$$

The corollary follows. □

For any $x^0 \in R^n$ and $K \subseteq N$, let

$$H(x^0, K) = \{x^0 + x \in R^n \mid 0 \leq x_i,\ i \in K, \text{ and } x_i = 0,\ i \notin K\},$$

and for any $x^0 \in R^n$, let $C(x^0)$ denote the closure of $R^n \backslash H(x^0, N)$. Then, as a direct result of Lemma 7, one can see that $H(x^{\max}, N) \cap P = x^{\max}$ and for any $x^0 \in R^n$ with $x^0 \geq x^{\max}$ and $x^0 \neq x^{\max}$, $H(x^0, N) \cap P = \emptyset$. These sets will play an important role in the following discussions.

LEMMA 9. *Both $C(x^0) \times \{1\}$ and $C(x^0) \times \{0\}$ contain a finite number of almost complete n-dimensional simplices of $\mathcal{D}_1$ that carry only integer labels in N.*

Proof. We will show only that $C(x^0) \times \{1\}$ contains a finite number of almost complete n-dimensional simplices of $\mathcal{D}_1$ that carry only integer labels in N. The result for $C(x^0) \times \{0\}$ can be obtained by replacing f with f^0 in the proof.

Let $\sigma = \langle (y^0, 1), (y^1, 1), \ldots, (y^n, 1) \rangle$ with $y^i \in R^n$, $i = 0, 1, \ldots, n$, be an arbitrary almost complete n-dimensional simplex of $\mathcal{D}_1$ that carries only integer labels in N. Without loss of generality, we assume that $l(y^i, 1) = i$, $i = 1, 2, \ldots, n$. From Definition 1, we know that, for $i = 1, 2, \ldots, n$, $f_i(y^i) > 0$ and

$$0 \geq f_j(y^i) - f_i(y^i), \tag{12}$$

$j = 1, 2, \ldots, n$. Let k be an index of N such that

$$0 \leq f_i(y^i) - f_k(y^k), \tag{13}$$

$i = 1, 2, \ldots, n$. Let $(x, 1)$ be an arbitrary point of σ. Then, as a result of (12), (13), Lemma 3, and $\text{mesh}(D_1) = 1$, we have, for $i = 1, 2, \ldots, n$,

$$\begin{aligned} f_i(x) - f_k(y^k) &= f_i(x) - f_i(y^k) + f_i(y^k) - f_k(y^k) \\ &\leq f_i(x) - f_i(y^k) \\ &\leq m\|x - y^k\| \\ &\leq m \end{aligned}$$

and

$$\begin{aligned} f_i(x) - f_k(y^k) &= f_i(x) - f_i(y^i) + f_i(y^i) - f_k(y^k) \\ &\geq f_i(x) - f_i(y^i) \\ &\geq -m\|x - y^i\| \\ &\geq -m. \end{aligned}$$

Thus,

$$\|f(x) - f_k(y^k)e\| \leq m,$$

where $e = (1, 1, \ldots, 1)^\top \in R^n$. Therefore, σ is contained in $W \times \{1\}$ with

$$W = \{x \mid \|f(x) - \mu e\| \leq m \text{ for some } \mu > 0\}.$$

Let x be an arbitrary point of W satisfying $f(x) > 0$. Then, from Corollary 2, we know that $x > y$ for any $y \in P$. Thus, $a_j^\top x \leq b_j$ for any $j \in M$ with $a_j \leq 0$. Let $J_i(x) = \{j \in J(x) \mid a_{ji} > 0\}$, $i = 1, 2, \ldots, n$. Then, $J_i(x) \neq \emptyset$, $i = 1, 2, \ldots, n$, and $J(x) = \cup_{i \in N} J_i(x)$. For $i = 1, 2, \ldots, n$, let j_i be any given index of $J_i(x)$ satisfying that

$$a_{j_i}^\top x - b_{j_i} = \max_{j \in J_i(x)} a_j^\top x - b_j.$$

Clearly, for any $j \in J_i(x)$, there exists uniquely $0 < r_j \leq 1$ satisfying that $a_j^\top x - b_j = r_j(a_{j_i}^\top x - b_{j_i})$. For $i = 1, 2, \ldots, n$, let

$$d_i = \sum_{j \in J_i(x)} \frac{r_j}{a_j^\top a_j} a_j = \frac{1}{a_{j_i}^\top a_{j_i}} a_{j_i} + \sum_{j \in J_i(x) \ \& \ j \neq j_i} \frac{r_j}{a_j^\top a_j} a_j.$$

Let $D = (d_1, d_2, \ldots, d_n)$, $\bar{A}^\top = (a_{j_1}, a_{j_2}, \ldots, a_{j_n})$, and $\bar{b} = (b_{j_1}, b_{j_2}, \ldots, b_{j_n})^\top$. Then,

$$\bar{A}x - \bar{b} = (a_{j_1}^\top x - b_{j_1}, a_{j_2}^\top x - b_{j_2}, \ldots, a_{j_n}^\top x - b_{j_n})^\top > 0$$

and

$$\begin{aligned} 0 < f(x) &= \textstyle\sum_{j \in J(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j \\ &= \textstyle\sum_{i=1}^n \sum_{j \in J_i(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j \\ &= \textstyle\sum_{i=1}^n \Big(\sum_{j \in J_i(x)} \frac{r_j}{a_j^\top a_j} a_j \Big)(a_{j_i}^\top x - b_{j_i}) \\ &= \textstyle\sum_{i=1}^n (a_{j_i}^\top x - b_{j_i}) d_i \\ &= D(a_{j_1}^\top x - b_{j_1}, a_{j_2}^\top x - b_{j_2}, \ldots, a_{j_n}^\top x - b_{j_n})^\top \\ &= D(\bar{A}x - \bar{b}) \\ &= D\bar{A}x - D\bar{b}. \end{aligned} \tag{14}$$

For $i = 1, 2, \ldots, n$, let

$$\bar{d_i} = \frac{1}{a_{j_i}^\top a_{j_i}} a_{j_i} + \sum_{j \in J_i(x) \ \& \ j \neq j_i} \frac{t_j}{a_j^\top a_j} a_j,$$

with t_j being an arbitrary number in $[0,1]$ for any $j \in J_i(x)$ and $j \neq j_i$. Let $\bar{D} = (\bar{d}_1, \bar{d}_2, \ldots, \bar{d}_n)$. Clearly, $\bar{D} = D$ when $t_j = r_j$ for any $j \in J_i(x)$ and $j \neq j_i$, $i = 1, 2, \ldots, n$. For any $i \in N$ and $x \in W$ with $f(x) > 0$, we have

$$
\begin{aligned}
\bar{d}_i^{\top} x &= \frac{a_{j_i}^{\top} x}{a_{j_i}^{\top} a_{j_i}} + \sum_{j \in J_i(x) \ \& \ j \neq j_i} \frac{t_j a_j^{\top} x}{a_j^{\top} a_j} \\
&= \frac{a_{j_i}^{\top} x}{a_{j_i}^{\top} a_{j_i}} + \sum_{j \in J_i(x) \ \& \ j \neq j_i} \frac{t_j b_j}{a_j^{\top} a_j} + \sum_{j \in J_i(x) \ \& \ j \neq j_i} \frac{t_j (a_j^{\top} x - b_j)}{a_j^{\top} a_j} \\
&\geq \frac{a_{j_i}^{\top} x}{a_{j_i}^{\top} a_{j_i}} + \sum_{j \in J_i(x) \ \& \ j \neq j_i} \frac{t_j b_j}{a_j^{\top} a_j}.
\end{aligned} \tag{15}
$$

From the definitions of $\bar{D}$, we know that $\bar{D}$ has exactly one positive entry in each row and each column and is bounded on $\{x \mid 0 < f(x)\}$. Therefore, by (14) and (15), we obtain that there exists a sufficiently large positive number δ satisfying that both $\bar{A}x > 0$ and $x^{\top}\bar{D} > 0$ whenever $f(x) \geq \delta e$.

Let

$$W^- = \{x \mid \|f(x) - \mu e\| \leq m \text{ for some } \mu \text{ with } 0 < \mu \leq m + \delta + 1\}$$

and

$$W^+ = \{x \mid \|f(x) - \mu e\| \leq m \text{ for some } \mu \text{ with } m + \delta + 1 \leq \mu\}.$$

Then, $W = W^- \cup W^+$. As follows, we show that both $C(x^0) \cap W^-$ and $C(x^0) \cap W^+$ are bounded.

From W, one can see that, for any $x \in W$,

$$-m|x_i| \leq x_i f_i(x) - \mu x_i \leq m|x_i|, \ i = 1, 2, \ldots, n.$$

Thus, for any $x \in W$,

$$-m \leq \frac{x^{\top} f(x)}{\|x\|_1} - \mu \frac{e^{\top} x}{\|x\|_1} \leq m.$$

By Lemma 4, we derive that W^- is bounded. Therefore, $C(x^0) \cap W^-$ is bounded.

Consider $C(x^0) \cap W^+$. Let x be an arbitrary point of W^+. Then,

$$(\delta + 1)e \leq (\mu - m)e \leq f(x) \leq (\mu + m)e.$$

From Lemma 8, we obtain that $\bar{A}$ and $\bar{D}$ are nonsingular and $\bar{A}^{-1} \geq 0$ and $\bar{D}^{-1} \geq 0$. From the definition of $\bar{D}$, we obtain that $\bar{D}^{-1}$ is bounded on W^+. Since $\bar{D} = D$ when $t_j = r_j$ for any $j \in J_i(x)$ and $j \neq j_i$, $i = 1, 2, \ldots, n$, hence, D is nonsingular and D^{-1} is bounded on W^+. From (14), we obtain that

$$x = \bar{A}^{-1} D^{-1} f(x) + \bar{A}^{-1} \bar{b}.$$

Since $(\mu - m)e \leq f(x) \leq (\mu - m)e$ and $\bar{A}^{-1} D^{-1} \geq 0$, hence,

$$(\mu - m)\bar{A}^{-1} D^{-1} e \leq x - \bar{A}^{-1} \bar{b} \leq (\mu + m)\bar{A}^{-1} D^{-1} e. \tag{16}$$

From (16), $\bar{A}^{-1} D^{-1} e > 0$, and the boundedness of D^{-1} on W^+, we derive that $C(x^0) \cap W^+$ is bounded. The lemma follows. □

LEMMA 10. *If $f(x) \leq 0$ and $f(x) \neq 0$, then, for any $x^0 \in P$, there is some k satisfying $x_k - x_k^0 < 0$.*

Proof. Suppose that $x - x^0 \geq 0$. Then,

$$\begin{aligned} 0 \geq (x - x^0)^\top f(x) &= \textstyle\sum_{j \in J(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} a_j^\top (x - x^0) \\ &= \textstyle\sum_{j \in J(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} (a_j^\top x - b_j + b_j - a_j^\top x^0) \\ &\geq \textstyle\sum_{j \in J(x)} \frac{a_j^\top x - b_j}{a_j^\top a_j} (a_j^\top x - b_j) > 0. \end{aligned}$$

A contradiction occurs. The lemma follows. ∎

LEMMA 11. *If z^0 is an integer point of P, then, for any $K \subseteq N$, $(x, 1) \in H(z^0, K) \times \{1\}$ carries a label of either 0 or an integer in K.*

Proof. From Lemma 10, we derive that no point of $H(z^0, K) \times \{1\}$ carries integer label $n + 1$. For $x \in H(z^0, K)$, let $\lambda = x - z^0$. Then, $0 \leq \lambda_j$, $j \in K$, and $\lambda_j = 0$, $j \notin K$. Thus, for every index i with $a_{ij} \leq 0$ for any $j \in K$,

$$\begin{aligned} a_i^\top x &= a_i^\top z^0 + a_i^\top \lambda \\ &\leq b_i + a_i^\top \lambda \\ &= b_i + \textstyle\sum_{j \in K} a_{ij} \lambda_j \\ &\leq b_i. \end{aligned}$$

Therefore, by Definition 1, we obtain that no point in $H(z^0, K) \times \{1\}$ carries an integer label in $N_0 \backslash K$. The lemma follows. ∎

As a corollary of Lemma 11, we have the following result.

COROLLARY 3. *If z^0 is an integer point of P, then there is no complete n-dimensional simplex in $H(z^0, N) \times \{1\}$ and there is no complete $(n-1)$-dimensional simplex in $\cup_{j \in N} H(z^0, N \backslash \{j\}) \times \{1\}$ that carries all integer labels in N.*

LEMMA 12. *For any subset K of N, $(x, 0) \in H(\eta, K) \times \{0\}$ with $x \neq \eta$ carries an integer label in K.*

Proof. For $x \in H(\eta, K)$, let $\lambda = x - \eta$. Then, $0 \leq \lambda_j$, $j \in K$, and $\lambda_j = 0$, $j \notin K$. Thus, for every index i with $a_{ij} \leq 0$ for any $j \in K$,

$$\begin{aligned} a_i^\top x &= a_i^\top \eta + a_i^\top \lambda \\ &= d_i + a_i^\top \lambda \\ &= d_i + \textstyle\sum_{j \in K} a_{ij} \lambda_j \\ &\leq d_i. \end{aligned}$$

Therefore, from Definition 1, we obtain that no point in $H(\eta, K) \times \{0\}$ with the exception of $(\eta, 0)$ carries an integer label in $N_0 \backslash K$. The lemma follows. ∎

As a corollary of Lemma 12, we have the following.

COROLLARY 4. *There is no complete $(n-1)$-dimensional simplex that is contained in $\cup_{j \in N} H(\eta, N \backslash \{j\}) \times \{0\}$ and carries all integer labels in N.*

3. An arbitrary starting homotopy-like simplicial algorithm. For any number α, let $\lfloor \alpha \rfloor$ denote the greatest integer less than or equal to α. We define $x^u = (x_1^u, x_2^u, \ldots, x_n^u)^\top$ with $x_i^u = \lfloor x_i^{\max} \rfloor$, $i = 1, 2, \ldots, n$. Then, $\lfloor x \rfloor \leq x^u$ for any $x \in P$.

LEMMA 13. *Let* $y = (\eta, 0)$, $\pi = (1, 2, \ldots, n+1)$, $s = (1, 1, \ldots, 1)^\top \in R^{n+1}$, $p = 0$, *and* $\sigma_0 = D_1(y, \pi, s, p) = \langle y^0, y^1, \ldots, y^{n+1} \rangle$ *with* $y^0 = y$ *and* $y^k = y + u^k$, $k = 1, 2, \ldots, n+1$. *Let* τ_0 *be the facet of* σ_0 *opposite to the vertex* y^{n+1}. *Then,* τ_0 *is a unique complete* n*-dimensional simplex that is contained in* $H(\eta, N) \times \{0\}$.

Proof. From Definition 1, we obtain that $l(y^0) = n+1$ and $l(y^k) = k$, $k = 1, 2, \ldots, n$. Thus, τ_0 is a complete n-dimensional simplex that is contained in $H(\eta, N) \times \{0\}$. From Lemma 12, we know that no point of $H(\eta, N) \times \{0\}$ carries integer label $n+1$ with the exception of $(\eta, 0)$. From the definition of the D_1-triangulation, one can see that τ_0 is a unique n-dimensional simplex in $\mathcal{D}_1$ that is contained in $H(\eta, N) \times \{0\}$ and has $(\eta, 0)$ as a vertex. The lemma follows. □

Based on the integer labeling rule in Definition 1 and the D_1-triangulation of $R^n \times [0, 1]$, an arbitrary starting homotopy-like simplicial algorithm is developed for computing an integer point in P of (1), which is as follows.

Initialization. Let $y = (\eta, 0)$, $\pi = (1, 2, \ldots, n+1)$, $s = (1, 1, \ldots, 1)^\top \in R^{n+1}$, and $p = 0$. Then, $\sigma_0 = D_1(y, \pi, s, p) = \langle y^0, y^1, \ldots, y^{n+1} \rangle$ is a unique $(n+1)$-dimensional simplex in $R^n \times [0, 1]$ having τ_0 as a facet. Let y^+ be the vertex of σ_0 opposite to τ_0, $q = 1$, and $k = 0$. Go to Step 1.

Step 1. Compute $l(y^+)$. If $l(y^+) = 0$, then the algorithm terminates and an integer point of P has been found. Otherwise, let y^- be the vertex of σ_k other than y^+ and carrying integer label $l(y^+)$, and τ_{k+1} the facet of σ_k opposite to y^-. Go to Step 2.

Step 2. If $\tau_{k+1} \subset R^n \times \{t\}$ for some $t \in \{0, 1\}$, go to Step 3. Otherwise, let σ_{k+1} be the unique $(n+1)$-dimensional simplex that is adjacent to σ_k and has τ_{k+1} as a facet, y^+ the vertex of σ_{k+1} opposite to τ_{k+1}, and $k = k+1$, and go to Step 1.

Step 3. If q is odd, then let $q = q+1$, $\sigma_{k+1} = \tau_{k+1}$, y^- be the vertex of σ_{k+1} carrying integer label $n+1$, τ_{k+2} the facet of σ_{k+1} opposite to y^-, and $k = k+1$, and go to Step 4. If q is even, then let $q = q+1$, $\tau_{k+1} = \sigma_k$, σ_{k+1} be the unique $(n+1)$-dimensional simplex in $R^n \times [0, 1]$ having τ_{k+1} as a facet, y^+ the vertex of σ_{k+1} opposite to τ_{k+1}, and $k = k+1$, and go to Step 1.

Step 4. Let σ_{k+1} be the unique n-dimensional simplex in $R^n \times \{t\}$ that is adjacent to σ_k and has τ_{k+1} as a facet, y^+ the vertex of σ_{k+1} opposite to τ_{k+1}, and $k = k+1$. Go to Step 5.

Step 5. Compute $l(y^+)$. If $l(y^+) = 0$, then the algorithm terminates and an integer point of P has been found. If $(x^u, 1) \leq y^+$, then the algorithm terminates and there is no integer point in P. If $l(y^+) = n+1$, then go to Step 3. If $l(y^+) \neq n+1$, then let y^- be the vertex of σ_k other than y^+ and carrying integer label $l(y^+)$, and τ_{k+1} the facet of σ_k opposite to y^-, and go to Step 4.

We remark that the algorithm consists of two phases. Steps 1–2 form one phase of the algorithm and Steps 4–5 form the other. Step 3 plays a bridge role for the algorithm to interchange from one phase to the other. As one may observe, the phase of Steps 1–2 comes from the well-known homotopy simplicial algorithm in [7, 9, 21] for computing fixed points.

THEOREM 2. *Within a finite number of iterations, the algorithm either yields an integer point of* P *or proves that no such point exists.*

To prove Theorem 2, we need to show first that the algorithm does not cycle. To accomplish this task, we will rely on an undirected graph. The way of defining the graph is similar in some aspects to that in [29].

For convenience of our further discussions, we introduce several shorthand notations, which are as follows:

1. for any $t \in \{0,1\}$, a CnS(t) stands for a complete n-dimensional simplex that is contained in $R^n \times \{t\}$;
2. a ZCnS stands for a 0-complete n-dimensional simplex that is contained in $R^n \times \{1\}$ and carries all integer labels in N;
3. for any $t \in \{0,1\}$, an ACnS(t) stands for an almost complete n-dimensional simplex that is contained in $R^n \times \{t\}$ and carries only integer labels in N;
4. a ZC$(n+1)$S stands for a 0-complete $(n+1)$-dimensional simplex; and
5. an AC$(n+1)$S stands for an almost complete $(n+1)$-dimensional simplex.

Let G be a graph given as follows.

- Nodes of G consist of
  1. all CnS(0)'s,
  2. all CnS(1)'s,
  3. all ZCnS's,
  4. all ACnS(0)'s,
  5. all ACnS(1)'s,
  6. all ZC$(n+1)$S's, and
  7. all AC$(n+1)$S's.
- There is an edge between two nodes of G if one is a complete n-dimensional facet of the other or they have either a common complete $(n-1)$-dimensional facet that carries all integer labels in N, or a common complete n-dimensional facet.

We say two nodes of a graph are adjacent if there is an edge between them. The degree of a node of a graph is equal to the number of nodes adjacent to it.

From the algorithm, one can see that when the phase of Steps 1–2 is executed, all the simplices generated by the algorithm are AC$(n+1)$S's before either a ZC$(n+1)$S or a CnS(t) for some $t \in \{0,1\}$ is met, and that for any given $t \in \{0,1\}$, when the phase of Steps 4–5 is executed, all the simplices generated by the algorithm are ACnS(t)'s before one of a ZCnS, a CnS(t), or a vertex $y^+ \in R^n \times \{t\}$ with $y^+ \geq (x^u, 1)$ is met. Thus, each simplex generated by the algorithm is exactly one of a CnS(0), a CnS(1), a ZCnS, an ACnS(0), an ACnS(1), a ZC$(n+1)$S, or an AC$(n+1)$S. Therefore, every simplex generated by the algorithm is a node of G and the algorithm moves from one node to one of its adjacent nodes along a connected component of G. To characterize the structure of a connected component of G, we need to determine the degree of each node of G, which is as follows.

1. Consider node σ of G given by a CnS(0). One can observe that node σ is adjacent only to a pair of nodes given by one of the following four pairs:
   (a) a CnS(0) and a ZC$(n+1)$S,
   (b) a CnS(0) and an AC$(n+1)$S,
   (c) an ACnS(0) and a ZC$(n+1)$S, or
   (d) an ACnS(0) and an AC$(n+1)$S.

   Thus, node σ has degree two. This is illustrated in Figure 1.
2. Consider node σ of G given by a CnS(1). One can observe that node σ is adjacent only to a pair of nodes given by one of the following three pairs:
   (a) a CnS(1) and an AC$(n+1)$S,
   (b) a ZCnS and an AC$(n+1)$S, or
   (c) an ACnS(1) and an AC$(n+1)$S.

   Thus, node σ has degree two. This is illustrated in Figure 2.

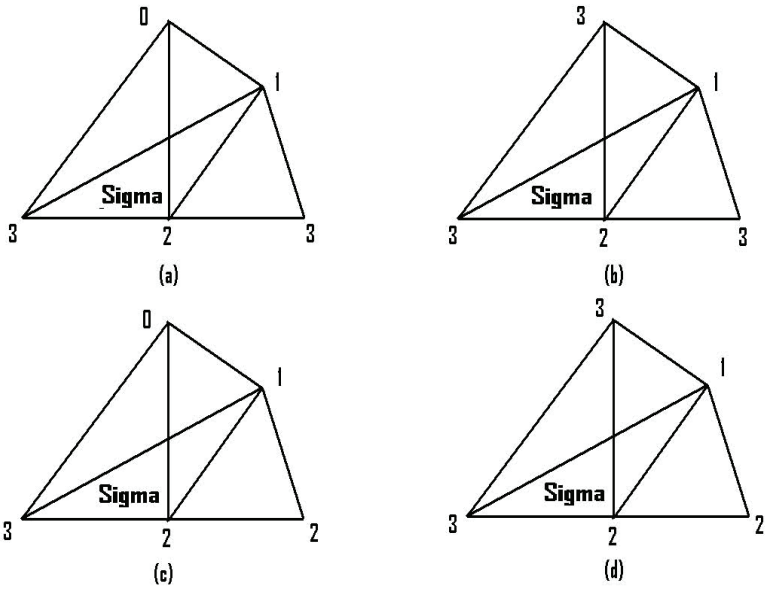


FIG. 1. *Illustration of the degree of node σ given by a $CnS(0)$ for $n = 2$.*

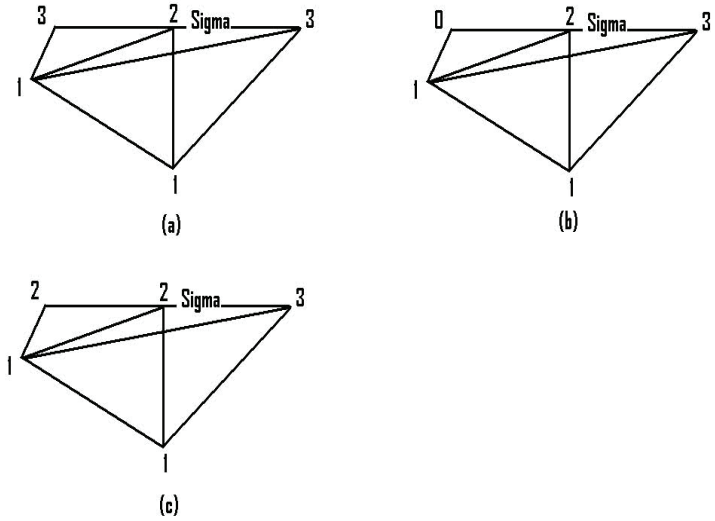


FIG. 2. *Illustration of the degree of node σ given by a $CnS(1)$ for $n = 2$.*

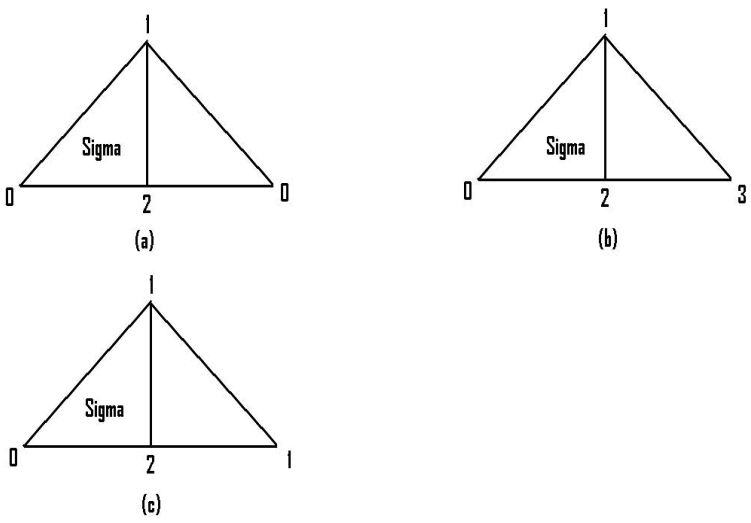


FIG. 3. *Illustration of the degree of node σ given by a ZCnS for $n = 2$.*

3. Consider node σ of G given by a ZCnS. One can observe that node σ is adjacent only to the node given by one of
   (a) a ZCnS,
   (b) a CnS(1), or
   (c) an ACnS(1).
   Thus, node σ has degree one. This is illustrated in Figure 3.
4. Consider node σ of G given by an ACnS(0). One can observe that node σ is adjacent only to a pair of nodes given by one of the following three pairs:
   (a) two CnS(0)'s,
   (b) a CnS(0) and an ACnS(0), or
   (c) two ACnS(0)'s.
   Thus, node σ has degree two. This is illustrated in Figure 4.
5. Consider node σ of G given by an ACnS(1). One can observe that node σ is adjacent only to a pair of nodes given by one of the following six pairs:
   (a) two ZCnS's,
   (b) a ZCnS and a CnS(1),
   (c) a ZCnS and an ACnS(1),
   (d) two CnS(1)'s,
   (e) a CnS(1) and an ACnS(1), or
   (f) two ACnS(1)'s.
   Thus, node σ has degree two. This is illustrated in Figure 5.
6. Consider node σ of G given by a ZC$(n+1)$S. One can observe that node σ is adjacent only to the node given by one of
   (a) a ZC$(n+1)$S,
   (b) a CnS(0), or
   (c) an AC$(n+1)$S.
   Thus, node σ has degree one. This is illustrated in Figure 6.

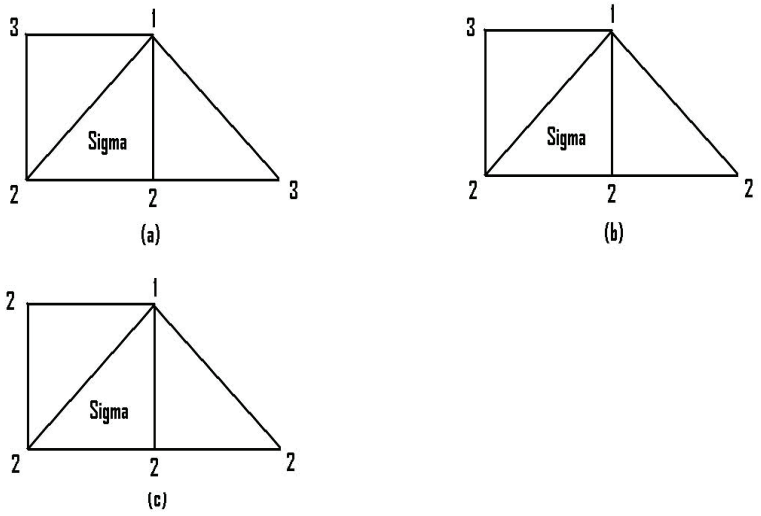


FIG. 4. *Illustration of the degree of node σ given by an $ACnS(0)$ for $n = 2$.*

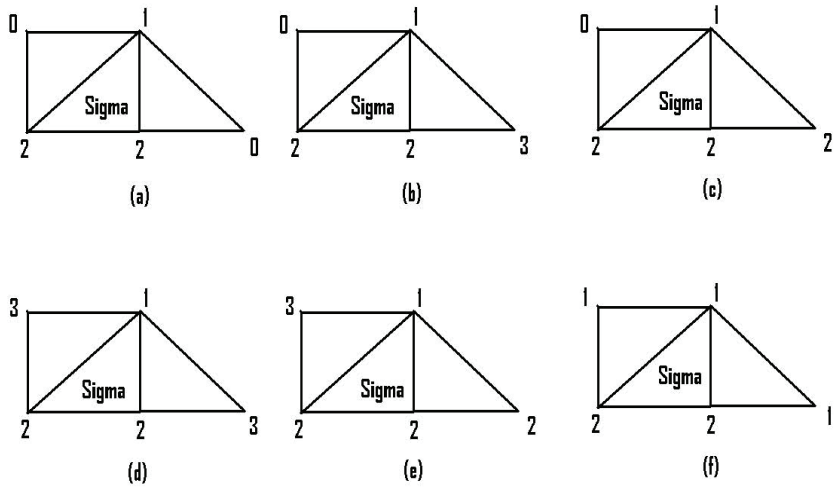


FIG. 5. *Illustration of the degree of node σ given by an $ACnS(1)$ for $n = 2$.*

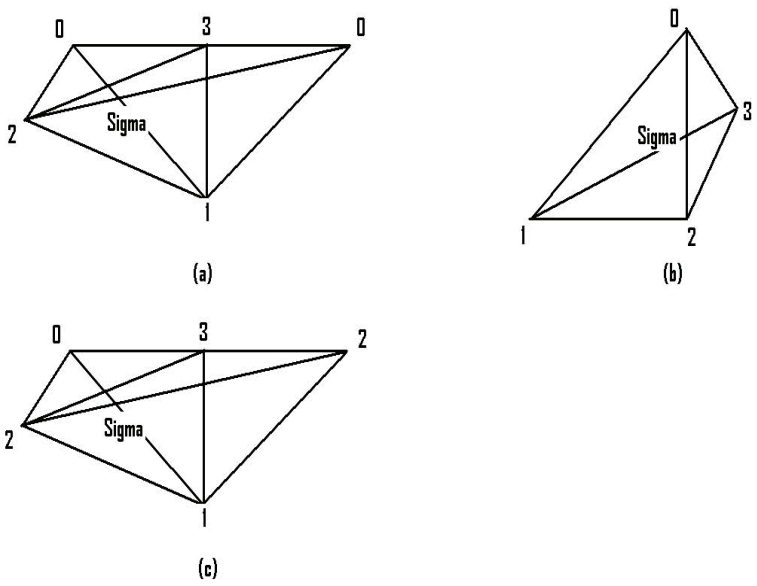


FIG. 6. *Illustration of the degree of node σ given by a ZC$(n+1)$S for $n=2$.*

7. Consider node σ of G given by an AC$(n+1)$S. One can observe that node σ is adjacent only to a pair of nodes given by one of the following five pairs:
   (a) two ZC$(n+1)$S's,
   (b) a ZC$(n+1)$S and a CnS(t) for some $t \in \{0,1\}$,
   (c) a ZC$(n+1)$S and an AC$(n+1)$S,
   (d) an AC$(n+1)$S and a CnS(t) for some $t \in \{0,1\}$, or
   (e) two AC$(n+1)$S's.

   Thus, node σ has degree two. This is illustrated in Figure 7.

   From the definition of G, one can see that each node of G belongs uniquely to one of these seven categories. The above results show that the degree of each node of G is at most two. Therefore, we come to the following conclusions.

   LEMMA 14. *Each connected component of graph G has one of the following forms:*
   - *A simple circuit, in which each of its nodes has degree two.*
   - *A simple path, in which each of its end nodes (if it has any) has degree one and is given by either a ZCnS or a ZC$(n+1)$S.*

   Consider the starting simplex of the algorithm, τ_0, given in Lemma 13. Since τ_0 is a CnS(0), hence, τ_0 is a node of G with degree two and there is a unique connected component of G that has τ_0 as a node. As a result of Corollary 4, one can obtain that the complete facet of τ_0 carrying all integer labels in N is not contained in $\cup_{j\in N} H(\eta, N\backslash\{j\}) \times \{0\}$. Thus, the pair of nodes of G adjacent to τ_0 is given by one of the following two pairs:
   - a ZC$(n+1)$S and an ACnS(0) contained in $H(\eta,N)\times\{0\}$, or
   - an AC$(n+1)$S and an ACnS(0) contained in $H(\eta,N)\times\{0\}$.

   This is illustrated in Figure 8.

Let P_{τ_0} be the unique connected component of G that has τ_0 as a node. Consider nodes of P_{τ_0} that are contained in $H(\eta,N)\times\{0\}$. From Lemma 13, we know that

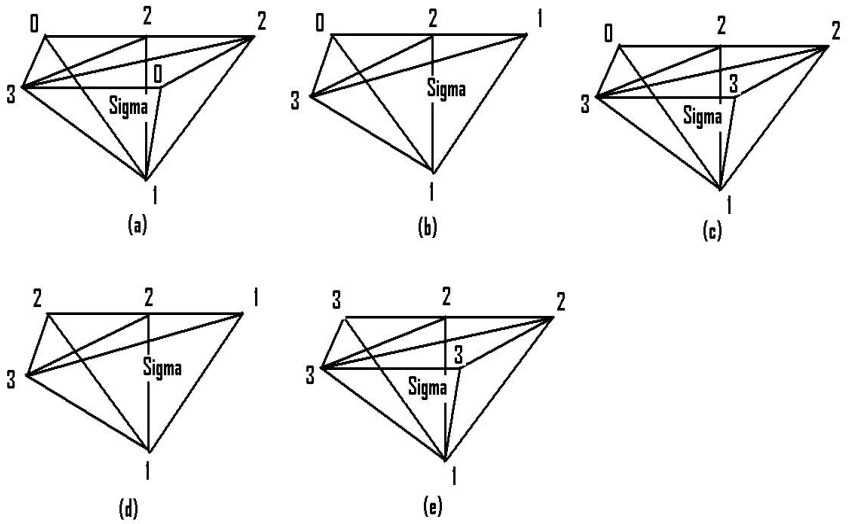


FIG. 7. *Illustration of the degree of node σ given by an $AC(n+1)S$ for $n = 2$.*

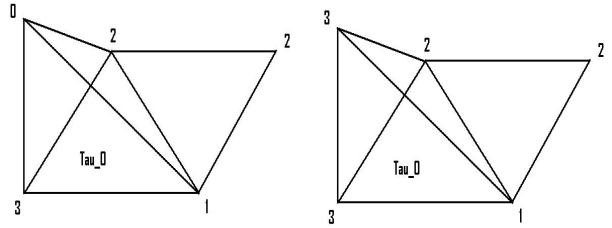


FIG. 8. *Illustration of the pair of nodes adjacent to τ_0 for $n = 2$.*

with the exception of τ_0, all the nodes of P_{τ_0} that are contained in $H(\eta, N) \times \{0\}$ are ACnS(0)'s. Applying Corollary 4, we derive that with the exception of τ_0, not one of the nodes of P_{τ_0} that are contained in $H(\eta, N) \times \{0\}$ is connected to any node that is not contained in $H(\eta, N) \times \{0\}$. Thus, all the nodes of P_{τ_0} that are contained in $H(\eta, N) \times \{0\}$ form an infinite simple path of G. Therefore, P_{τ_0} is an infinite simple path of G.

Let $P^0_{\tau_0}$ be the part of P_{τ_0} that starts from τ_0 and has no nodes in $H(\eta, N) \times \{0\}$ with the exception of τ_0. Then, one can see that the algorithm exactly follows the path $P^0_{\tau_0}$ to move from one node to another before it terminates. Therefore, we come to the conclusion that the algorithm does not cycle.

Proof of Theorem 2. From the above results, we know that the algorithm does not cycle. Thus, all the simplices generated by the algorithm are different from each other.

1. Assume that the phase of Steps 1–2 is executed. From the algorithm, we know that before the interchanging of two phases, all the simplices generated

by the algorithm in an execution of the phase of Steps 1–2 are AC$(n+1)$S's. Thus, as a result of Corollary 1, we obtain that, within a finite number of iterations, an execution of the phase of Steps 1–2 terminates with either an integer point of P or a CnS(t) for some $t \in \{0,1\}$. If an execution of the phase of Steps 1–2 generates a CnS(t) for some $t \in \{0,1\}$, then the algorithm interchanges from the phase of Steps 1–2 to the phase of Steps 4–5.

2. Assume that the phase of Steps 4–5 is executed. Let σ be the starting simplex, which must be a CnS(t) for some $t \in \{0,1\}$ generated by the algorithm in the phase of Steps 1–2.
Consider the case of $t = 0$. From the algorithm, one can see that all the simplices generated by the algorithm in the phase of Steps 4–5 are ACnS(0)'s before a CnS(0) is obtained. By Lemma 13, we know that σ is contained in $C(\eta) \times \{0\}$. Applying Corollary 4, we derive that all the ACnS(0)'s generated by the algorithm in the phase of Steps 4–5 are contained in $C(\eta) \times \{0\}$. As a result of Lemma 9, we obtain that, within a finite number of iterations, an execution of Steps 4–5 terminates with a CnS(0), and the algorithm interchanges from the phase of Steps 4–5 to the phase of Steps 1–2.
Consider the case of $t = 1$.
   - Assume that P has an integer point. Let z^0 be an integer point of P. Then, $z^0 \leq x^u$. Applying Corollary 3, we obtain that σ and all the ACnS(1)'s generated by the algorithm in the phase of Steps 4–5 are contained in $C(z^0) \times \{1\}$. As a result of Lemma 9, we derive that, within a finite number of iterations, an execution of the phase of Steps 4–5 terminates with either an integer point of P or a CnS(1). If an execution of the phase of Steps 4–5 generates a CnS(1), then the algorithm interchanges from the phase of Steps 4–5 to the phase of Steps 1–2.
   - Assume that P has no integer point. From Step 5 of the algorithm, one can see that all the ACnS(1)'s generated by the algorithm in the phase of Steps 4–5 are contained in $C(x^u) \times \{1\}$. As a result of Lemma 9, we obtain that, within a finite number of iterations, an execution of the phase of Steps 4–5 terminates with either a CnS(1) or a point $y \in R^n \times \{1\}$ such that $y \geq (x^u, 1)^\top$. If an execution of the phase of Steps 4–5 generates a CnS(1), then the algorithm interchanges from the phase of Steps 4–5 to the phase of Steps 1–2.

   As a result of Theorem 1, we know that the algorithm interchanges between two phases at most a finite number of times. Therefore, combining the above results, we come to the conclusion that, within a finite number of iterations, the algorithm either yields an integer point of P or proves no such point exists. The theorem follows. □

REFERENCES

[1] E. L. Allgower and K. Georg, *Piecewise linear methods for nonlinear equations and optimization*, J. Comput. Appl. Math., 124 (2000), pp. 245–261.
[2] C. Dang, *The D_1-triangulation of R^n for simplicial algorithms for computing solutions of nonlinear equations*, Math. Oper. Res., 16 (1991), pp. 148–161.
[3] C. Dang, *Triangulations and Simplicial Methods*, Lecture Notes in Econom. and Math. Systems 421, Springer-Verlag, Berlin, 1995.
[4] C. Dang and H. van Maaren, *A simplicial approach to the determination of an integral point of a simplex*, Math. Oper. Res., 23 (1998), pp. 403–415.
[5] C. Dang and H. van Maaren, *An arbitrary starting variable dimension algorithm for com-*

puting an integer point of a simplex, Comput. Optim. Appl., 14 (1999), pp. 133–155.
[6] C. Dang and H. van Maaren, *Computing an integer point of a simplex with an arbitrary starting homotopy-like simplicial algorithm*, J. Comput. Appl. Math., 129 (2001), pp. 151–170.
[7] B. C. Eaves, *Homotopies for the computation of fixed points*, Math. Program., 3 (1972), pp. 1–22.
[8] B. C. Eaves, *A Course in Triangulations for Solving Equations with Deformations*, Lecture Notes in Econom. and Math. Systems 234, Springer-Verlag, Berlin, 1984.
[9] B. C. Eaves and R. Saigal, *Homotopies for the computation of fixed points on unbounded regions*, Math. Program., 3 (1972), pp. 225–237.
[10] W. Forster, *Homotopy methods*, in Handbook of Global Optimization, R. Horst and P. M. Pardalos, eds., Kluwer Academic Publishers, 1995, pp. 669–750.
[11] H. Freudenthal, *Simplizialzerlegungen von beschränkter Flachheit*, Ann. of Math. (2), 43 (1942), pp. 580–582.
[12] C. B. Garcia and W. I. Zangwill, *Pathways to Solutions, Fixed Points, and Equilibria*, Series in Computational Mathematics, Prentice–Hall, Englewood Cliffs, NJ, 1981.
[13] D. S. Hochbaum and J. Naor, *Simple and fast algorithms for linear and integer programs with two variables per inequality*, SIAM J. Comput., 23 (1994), pp. 1179–1192.
[14] R. Kannan, *Polynomial-time aggregation of integer programming problems*, J. Assoc. Comput. Mach., 30 (1983), pp. 133–145.
[15] M. Kojima and Y. Yamamoto, *A unified approach to the implementation of several restart fixed point algorithms and a new variable dimension algorithm*, Math. Program., 28 (1984), pp. 288–328.
[16] H. W. Kuhn, *Simplicial approximation of fixed points*, Proc. Natl. Acad. Sci. USA, 61 (1968), pp. 1238–1242.
[17] G. van der Laan and A. J. J. Talman, *A restart algorithm for computing fixed points without an extra dimension*, Math. Program., 17 (1979), pp. 74–84.
[18] G. van der Laan and A. J. J. Talman, *A class of simplicial restart fixed point algorithms without an extra dimension*, Math. Program., 20 (1981), pp. 33–48.
[19] J. C. Lagarias, *The computational complexity of simultaneous Diophantine approximation problems*, SIAM J. Comput., 14 (1985), pp. 196–209.
[20] H.-J. Lüthi, *A simplicial approximation of a solution for the nonlinear complementarity problem*, Math. Program., 9 (1975), pp. 278–293.
[21] O. H. Merrill, *Applications and Extensions of an Algorithm that Computes Fixed Points of Certain Upper Semi-Continuous Point to Set Mappings*, Ph.D. Thesis, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, 1972.
[22] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*, Wiley, New York, 1998.
[23] A. Pnueli, *A Method of Truncated Relaxation for Integer Programming*, RC 2267, IBM Research, Research Division, Yorktown Heights, NY, 1968.
[24] H. Scarf, *The approximation of fixed points of a continuous mapping*, SIAM J. Appl. Math., 15 (1967), pp. 1328–1343.
[25] H. Scarf, *The Computation of Economic Equilibria*, Yale University Press, New Haven, 1973.
[26] H. E. Scarf, *Production sets with indivisibilities*. I. *Generalities*, Econometrica, 49 (1981), pp. 1–32.
[27] H. E. Scarf, *Neighborhood systems for production sets with indivisibilities*, Econometrica, 54 (1986), pp. 507–532.
[28] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley, New York, 1998.
[29] M. J. Todd, *The Computation of Fixed Points and Applications*, Lecture Notes in Econom. and Math. Systems 124, Springer-Verlag, Berlin, 1976.
[30] A. H. Wright, *The octahedral algorithm, a new simplicial fixed point algorithm*, Math. Program., 21 (1981), pp. 47–69.
[31] Y. Yamamoto, *A new variable dimension algorithm for the fixed point problem*, Math. Program., 25 (1983), pp. 329–342.

SIAM J. DISCRETE MATH.
Vol. 23, No. 2, pp. 634–646

PEBBLING ALGORITHMS IN DIAMETER TWO GRAPHS*

AIRAT BEKMETJEV† AND CHARLES A. CUSACK‡

Abstract. Consider a connected graph and a configuration of pebbles on its vertices. A pebbling step consists of removing two pebbles from a vertex and placing one on an adjacent vertex. A configuration is called solvable if it is possible to place a pebble on any given vertex through a sequence of pebbling steps. A smallest number t such that any configuration with t pebbles is solvable is called the pebbling number of the graph. In this paper, we consider algorithms determining the solvability of a pebbling configuration on graphs of diameter two. We prove that if k is the vertex connectivity of a diameter two graph G, then a configuration is solvable if there are at least $c(k) = \min\{k+4, 3k-1\}$ vertices in G with two or more pebbles. We use this result to construct an algorithm that has complexity $O(c(k)! \cdot n^{2c(k)-3}m)$, where n is the number of vertices and m is the number of edges. We also present an algorithm for diameter two graphs with pebbling number $n+1$, known as Class 1 graphs, which takes $O(nm)$ time.

Key words. graph pebbling, diameter, connectivity, algorithms

AMS subject classifications. 05C85, 05C40, 68R05

DOI. 10.1137/080724277

1. Introduction.

Let G be a connected graph with vertex set V and edge set E, with $n = |V|$ and $m = |E|$. Define a *pebbling configuration* as a function $C : V \to Z^+$, where $C(v)$ represents the number of pebbles placed on vertex v. For any vertex v with $C(v) \geqslant 2$, a *pebbling step* consists of placing one pebble on an adjacent vertex and discarding two pebbles from v. A configuration is called *r-solvable* if there is a sequence of pebbling steps that places at least one pebble on vertex r. Any such sequence is called an *r-solution*. A configuration is called *solvable* if it is r-solvable for any $r \in V$. We call an r-solution *minimal* if it contains the smallest number of pebbling steps.

This paper considers an algorithmic approach to the pebbling problem. Watson [8] and Milans and Clark [6] showed that determining the solvability of a pebbling configuration on a general graph is an NP-complete problem. We will consider graphs of diameter two and show the existence of an algorithm whose running time depends on the vertex connectivity and the size of the graph. In particular, we will show that in a diameter two graph with connectivity k, any configuration that contains at least $c(k) = \min\{3k-1, k+4\}$ vertices with two or more pebbles is solvable. Based on this result, we will establish an algorithm that determines the solvability of a given configuration in $O(c(k)! \cdot n^{2c(k)-3}m)$ time, which is polynomial when k is constant.

We begin by presenting a backtracking algorithm (Algorithm 1.1, which uses Algorithm 1.2) that determines the solvability of a pebbling configuration on any graph. The method AdjacentPebble(u, v) performs a pebbling step from u to v, assuming that u and v are adjacent, and that $C(u) \geqslant 2$. UndoPebble(u, v) reverses

*Received by the editors May 13, 2008; accepted for publication (in revised form) November 3, 2008; published electronically February 6, 2009. This work was supported in part by a grant to Hope College from the Howard Hughes Medical Institute through the Undergraduate Science Education Program.

http://www.siam.org/journals/sidma/23-2/72427.html

†Department of Mathematics, Hope College, 27 Graves Place, Holland, MI 49422 (bekmetjev@hope.edu).

‡Department of Computer Science, Hope College, 27 Graves Place, Holland, MI 49422 (cusack@hope.edu).

a pebbling move. The algorithm maintains a set L of vertices which can be pebbled, returning TRUE if $|L| = n$ (since then all of the vertices have been covered), and returning FALSE if $|L| < n$ at the end of the algorithm.

The algorithm is based on the following ideas. If C is a solvable configuration, and r is a vertex with no pebbles, it follows from [6] that there is an acyclic r-solution, and all moves from valid vertices (i.e., vertices that contain at least two pebbles) can be made in an arbitrary order. Also, if M is an r-solution, then any pebbling sequence N with $M \subseteq N$ is also an r-solution.

Algorithm 1.1. IsSolvable(G, C).

global Set L
for $u \leftarrow 0$ **to** $n-1$
  do { **if** $C(u) \geqslant 1$ **then** add u to L } 1.1.1
if $|L| = n$
  then return (TRUE)
  else return (IsSolvableRecursive(G, C))

Algorithm 1.2. IsSolvableRecursive(G, C).

comment: Determine first vertex with at least 2 pebbles

$u \leftarrow 0$
while $u < n$ **and** $C(u) \leqslant 1$
  do $u \leftarrow u + 1$
if $u = n$
  then return (FALSE) } 1.2.1

comment: Now try all possible moves from u

for each v adjacent to u
  do { add v to L
  if $|L| = n$
    then return (TRUE)
  C.AdjacentPebble(u, v)
  solvable = IsSolvableRecursive(G, C')
  C.UndoPebble(u, v)
  if *solvable*
    then return (TRUE) } 1.2.2
return (FALSE)

Let $T(t)$ be the worst-case time it takes to determine the solvability of a graph with t pebbles using Algorithm 1.1, and let d be the maximum degree of G. Notice that step 1.1.1 in Algorithm 1.1 and step 1.2.1 in Algorithm 1.2 each take $O(n)$ time. A graph with a single pebble will end at step 1.2.1, so $T(1) = O(n)$. Step 1.2.2 executes at most d times, each time requiring $O(1)$ time plus making a recursive call on a graph with one fewer pebble. Thus,

$$T(t) = d(T(t-1) + O(1)) + O(n) = d \cdot T(t-1) + O(n) = O(nd^{t-1}).$$

Since t may depend on n, Algorithm 1.1 is not polynomial time in general. One source of inefficiency in this algorithm is that there may be many ways of moving a pebble

from one vertex to another along paths of vertices which contain a single pebble, and it tries all of them. It turns out that, for graphs of diameter two with connectivity k, we can avoid such an exhaustive search.

2. Diameter and connectivity. Let $\mathcal{G}_2$ represent the set of all graphs of diameter two, and let $\mathcal{G}_{2,k} \subset \mathcal{G}_2$ be the set of diameter two graphs which have vertex connectivity k. It is clear that if Q is a vertex cut set in a diameter two graph, then any vertex in $V \backslash Q$ must be adjacent to at least one vertex in Q. This observation leads to the following.

LEMMA 2.1. *Let $G \in \mathcal{G}_2$ and let Q be a vertex cut set. Then a configuration C is solvable if it is possible to place at least two pebbles on each vertex in Q.*

Let $C_m = \{v \in V \mid C(v) \geqslant m\}$. Note that for a graph $G \in \mathcal{G}_2$, any configuration is solvable if C_4 is nonempty. We are going to establish two upper bounds on $|C_2|$ for members of $\mathcal{G}_{2,k}$ which are unsolvable.

LEMMA 2.2. *Let $G \in \mathcal{G}_{2,k}$. Then a configuration C is solvable if $|C_2| \geqslant 3k - 1$.*

Proof. Let Q be a minimal cut set of G that contains k vertices. For any sequence of pebbling moves in an unsolvable configuration, none of the vertices in Q can accumulate four or more pebbles and, by Lemma 2.1, at least one vertex in Q can accumulate at most one. Therefore, at most $3(k-1)+1 = 3k-2$ pebbles can be placed on vertices in Q without the configuration being solvable. However, at least $3k-1$ pebbles can be placed on Q from C_2, and therefore C is solvable. ∎

The last result is tight for $k = 1, 2$. If $G \in \mathcal{G}_{2,1}$, then it has a vertex u that is not adjacent to all other vertices, and placing two pebbles on u and zero on all other vertices creates an unsolvable configuration. Further, Figure 2.1 represents a graph in $\mathcal{G}_{2,2}$ and a configuration C with $|C_2| = 4$, which is not solvable. As it is shown in the following result, the upper bound on $|C_2|$ can be improved for $k \geqslant 3$.

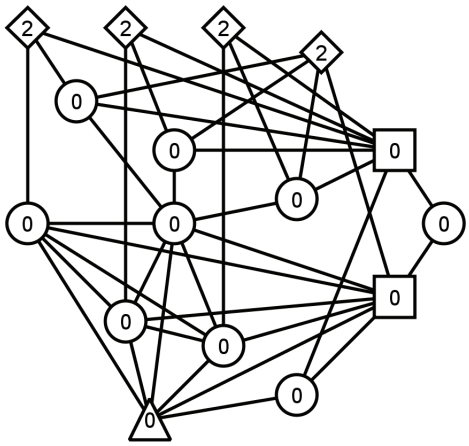


FIG. 2.1. *An unsolvable configuration for a graph from $\mathcal{G}_{2,2}$ with $|C_2| = 4$. The squares represent the cut set, the diamonds are the vertices containing two pebbles, and the triangle is the root.*

THEOREM 2.3. *Let $G \in \mathcal{G}_{2,k}$. Then a configuration C is solvable if $|C_2| \geqslant k+4$.*

Proof. Let C be an unsolvable configuration with $|C_2| = k + 4 + i$, where $i \geqslant 0$. Let Q be a vertex cut set of size k in G, $Q_2 = C_2 \cap Q$, and $Q_{0,1} = Q \setminus Q_2$. Also, let X be the set of vertices in the component of $V \setminus Q$ such that $|C_2 \cap X|$ is the smallest, and let $Y = V \setminus (Q \cup X)$. Let $X_2 = C_2 \cap X$, $Y_2 = C_2 \cap Y$, $|Q_2| = q$, and $|X_2| = x$.

Then $|Y_2| = k + 4 + i - q - x$. By construction, $|X_2| \leqslant |Y_2|$, so $x \leqslant \frac{k+4+i-q}{2}$. Finally, let $Q' \subseteq Q_{0,1}$ be the set of vertices that are adjacent to vertices in both X_2 and Y_2. By Lemma 2.1, at least one vertex in $Q_{0,1}$ must be adjacent to at most one vertex in $X_2 \cup Y_2$ (or the configuration will be solvable), so $|Q'| \leqslant k - q - 1$.

There are three cases to consider depending on the value of x.

1. $(x \geqslant 2)$. Let $v_2 \in X_2$ and $u_2 \in Y_2$. Any vertex $u \in Q$ that is adjacent to both v_2 and u_2 must be in $Q_{0,1}$, since otherwise two more pebbles can be placed on u from v_2 and u_2, giving it four. For the same reason, each $u \in Q_{0,1}$ can be adjacent to at most three vertices in $X_2 \cup Y_2$. This implies that each $u \in Q_{0,1}$ can be adjacent to at most two distinct pairs in $X_2 \times Y_2$. Therefore $|X_2||Y_2| = x(k + 4 + i - q - x) \leqslant 2(k - q - 1)$ or, equivalently,

$$(x - 2 + q - k - i)(x - 2) \geqslant 6 + 2i. \tag{2.1}$$

Since $x \leqslant \frac{k+4+i-q}{2}$, and $q \leqslant k$,

$$\begin{aligned}(x - 2 + q - k - i)(x - 2) &\leqslant \left(\frac{k + 4 + i - q}{2} - 2 + q - k - i\right)(x - 2)\\ &= \left(\frac{q - k - i}{2}\right)(x - 2)\\ &\leqslant 0,\end{aligned}$$

contradicting (2.1).

2. $(x = 1)$. Let $X = \{u\}$. Then there is a path of length two from each of the $k - q + 3 + i$ vertices in Y_2 to u with some vertex in Q' as the intermediate vertex. Thus, it is possible to move at least $k - q + 3 + i$ pebbles onto the vertices of Q'. Since $|Q'| \leqslant k - q - 1$, then $k - q + 3 + i \geqslant |Q'| + 4$, so either one of these vertices accumulates four or more, or at least two of them accumulates two or more. In either case, two pebbles can be moved onto u, giving it four pebbles.
3. $(x = 0)$. Let $u \in X$. Then there is a path of length two from each of the $k + 4 + i - q$ vertices in Y_2 to u with some vertex in Q as the intermediate vertex. Let $S \subseteq Q$ be the set of vertices that are adjacent to both u and a vertex in Y_2, and let $S_2 = S \cap Q_2$. Then $|S| \leq k - q + |S_2|$, and it is possible to accumulate at least $|Y_2| + 2|S_2| = k + 4 + i - q + 2|S_2| \geqslant |S| + 4 + |S_2| + i \geqslant |S| + 4$ pebbles onto the vertices in S. There are four cases to consider, each of which leads to a solvable configuration.
   (a) Some vertex $v \in S$ can accumulate four pebbles.
   (b) There are four vertices in S which can accumulate two pebbles, in which case four pebbles can be moved to u.
   (c) Some vertex $v_1 \in S$ can accumulate three pebbles, and $v_2, v_3 \in S$ can accumulate at least two each. Then two pebbles can be moved from v_2 and v_3 onto u, and then one can be moved from u to v_1, so that v_1 accumulates four.
   (d) $v_1, v_2 \in S$ can each accumulate three pebbles, and every other vertex in S can accumulate only one pebble. In this case, $S = Q_{0,1}$, and every vertex in $S \setminus \{v_1, v_2\}$ can accumulate exactly one pebble. Then one pebble can be moved from each of v_1 and v_2 onto u, and then one pebble can be moved from u onto any vertex in S. Thus, any vertex in Q can

accumulate at least two pebbles and, by Lemma 2.1, the configuration is solvable. □

COROLLARY 2.4. *Let $G \in \mathcal{G}_{2,k}$. Then a configuration C is solvable if $|C_2| \geqslant \min\{3k-1, k+4\}$.*

3. Islands and bridges. An *i-island* is a maximal connected subgraph of a graph in which every vertex has at least one pebble and one vertex has at least i pebbles. Given a set of integers $\mathbf{b} = \{b_1, b_2, \ldots, b_l\}$, a *$\mathbf{b}$-island* is an island that contains distinct vertices $v_1, v_2, \ldots, v_l$ such that $C(v_j) \geqslant b_j, 1 \leqslant j \leqslant l$. A *bridge* is a vertex which has zero pebbles. Notice that if a bridge v is adjacent to a 2-island, then one pebble can be added to v. If every vertex on an island contains precisely one pebble, we call it a *desert*. For an island I, the *surplus* $s_I(C)$ of I is the difference between the number of pebbles placed on I and the number of vertices in I, i.e.,

$$s_I(C) = \sum_{v \in I} C(v) - |I|.$$

We also define the surplus of the graph $s_G(C)$ as the sum of the surpluses of all islands. That is,

$$s_G(C) = \sum_{v \in G} C(v) - |C_1|.$$

Note that a pebbling move from an island to a bridge always reduces the surplus of the graph. In a graph with surplus s, for any vertex r, any r-solution can move pebbles onto at most s bridges, including r.

A vertex in G that is adjacent to at least one vertex of a subgraph H is called *adjacent to H*. Note that a vertex containing two or more pebbles allows the movement of at least one pebble to any other vertex on its island or any bridge adjacent to its island.

Every vertex of a graph is either a bridge or a member of a single island, and a pebbling configuration is solvable if and only if every bridge can be pebbled. In any sequence of pebbling steps, we say a bridge is *filled* if two pebbles are moved onto it, and *emptied* if two pebbles are removed from it. In a minimal r-solution for any root r, every bridge that is used must be filled and emptied, except r.

Recall that for any two vertices $u, v \in V$, the *distance* $d(u, v)$ between u and v is the number of edges on the shortest path connecting them. For a subset $S \subseteq V$, let $d_{\min}(v, S)$ be the smallest distance between vertex v and any vertex in S. Let

$$D_m(S) = \{v \in V \mid d_{\min}(v, S) = m\}.$$

In particular, for an island I, $D_1(I)$ is the set of vertices adjacent to at least one vertex in I.

LEMMA 3.1. *Let $G \in \mathcal{G}_2$ and let C be a configuration which contains an island I with $s_I(C) \geqslant 3$. Then C is solvable.*

Proof. If $s_I(C) \geqslant 3$, then I is a 4-island, a $\{2,3\}$-island, or a $\{2,2,2\}$-island. Clearly, a 4-island guarantees solvability. Given a $\{2,3\}$-island, we can move a pebble from the vertex with two pebbles to the vertex with three pebbles along a path in I, creating a solvable configuration. If I is a $\{2,2,2\}$-island, let a, b, and c be vertices in $C_2 \cap I$. Consider any path P between a and b. If $c \in P$, we can accumulate four pebbles on c from a and b. Otherwise, choose a shortest path P' between c and the vertices in P. Let $u = P \cap P'$. If $u = a$ (or $u = b$), we can accumulate four pebbles

on u by moving one from b (or a), and the other from c. If u is different from a and b, we can move three pebbles from a, b, and c onto u, giving it four pebbles. □

In light of Lemma 3.1, we call an island I with $s_I(C) \geqslant 3$ an *empire*. If a graph does not contain an empire, then the only possible islands are deserts, 2-islands, 3-islands, and $\{2,2\}$-islands. Note that the latter three cases of islands are not mutually exclusive.

Islands and pebbles can be maintained as part of the graph data structure, and each vertex can store a reference to the island to which it belongs. The following basic procedures are needed in the construction of an algorithm to determine the solvability of graphs in $\mathcal{G}_{2,k}$. An upper bound for their running time is provided. Notice that whenever a pebbling move is performed, the islands must be updated, since a single pebbling move might significantly change the configuration of islands. Thus UpdateIslands is called at the end of any procedure that moves pebbles.

- UpdateIslands() updates the data structure representing the islands and records whether or not an empire is present. This can be implemented using a standard BFS/DFS algorithm in $O(n+m)$ time (see [3], for instance).
- ContainsEmpire() returns TRUE if and only if G contains an empire. This takes $O(1)$ time.
- IsAdjacentTwoIsland(u) returns TRUE if and only if u is adjacent to a 2-island. This can be implemented in $O(d)$ time, where d is the maximum degree of G, by checking if any of the vertices adjacent to u belong to a 2-island.
- PebbleFromIsland(I, u) performs a sequence of pebbling steps required to move a pebble from I to u using only vertices in I in intermediate steps, assuming that $u \in D_1(S)$ and I is a 2-island. It calls UpdateIslands() and returns the resulting graph configuration. It takes $O(n+m)$ time.
- CanDoublePebble(G', I, u, v) returns TRUE if and only if pebbles can be moved from I (assuming that I is a $\{2,2\}$-island) to vertices u and v (where $u = v$ is possible) simultaneously, using only vertices from I in intermediate moves. It assigns to G' the resulting graph configuration and calls UpdateIslands() on G'. It takes $O(nm)$ time (see Corollary 3.3 below).

The next result by Shiloach [7] is used to determine the pebbling solvability in the presence of a $\{2,2\}$-island.

THEOREM 3.2 (see [7]). *For any distinct vertices s_1, s_2, t_1, and t_2, it can be determined in $O(nm)$ time whether or not G admits two vertex-disjoint paths connecting s_1 to t_1 and s_2 to t_2.*

COROLLARY 3.3. *Let I be a $\{2,2\}$-island that is not an empire.*

1. *For any two distinct vertices $u, v \in D_1(I)$, it can be determined in $O(nm)$ time whether or not both u and v can be pebbled from I simultaneously.*
2. *For any $u \in D_1(I)$, it can be determined in $O(nm)$ time whether u can be filled from I.*

Proof. Let G' be the subgraph of G induced by the vertex set $I \cup u \cup v$, and s_1 and s_2 be elements of $C_2 \cap I$. Then the first condition follows from Theorem 3.2 by determining disjoint paths connecting either s_1 to u and s_2 to v, or s_1 to v and s_2 to u in G'.

For the second condition, notice that if u is adjacent only to one vertex in I, then it cannot be filled from I. Otherwise, vertex u can be filled from I if and only if there are disjoint paths to u from $\{s_1, s_2\} = C_2 \cap I$. Let G' be a graph induced by the vertex set $I \cup u$ with an added vertex u' that is adjacent to everything that u is. Then there are disjoint paths from s_1 to u and s_2 to u in G if and only if there are disjoint paths from s_1 to u and s_2 to u' in G'. The result follows from Theorem 3.2. □

4. Algorithms for $\mathcal{G}_{2,k}$. The goal of this section is to present a polynomial time algorithm to determine the solvability of a pebbling configuration for graphs in $\mathcal{G}_{2,k}$.

Let C be a configuration on $G \in \mathcal{G}_2$ and I an island in C. In any sequence of pebbling moves, if one or more pebbles are moved from the vertices of I, and no pebbles are moved onto I, then I is called an *origin*.

THEOREM 4.1. *Let C be a solvable configuration which does not contain an empire and contains some vertex r with $C(r) = 0$ that is not adjacent to any 2-island. Then, any minimal r-solution contains an origin.*

Proof. Let b be the number of bridges in C used in a minimal r-solution, counting repeated use of any bridge. Note that $b \geqslant 1$, since r is not adjacent to a 2-island. Further, in a minimal r-solution, exactly $2b$ pebbles are moved to the bridges, and exactly b pebbles are moved from the bridges to adjacent vertices. Let j be the number of these b pebbles that is moved onto bridges or deserts in C. Clearly, $j \geqslant 1$, since one pebble must be moved either to r or a desert adjacent to r. Thus, at most $b-j$ pebbles can be moved onto 2-islands, since 2-islands in C can be pebbled only from bridges in C. If there is no origin, the number of 2-islands used in the pebbling is at most $b-j$. Since there is no empire, the only possible islands in C are 2-islands (including 3-islands) and $\{2,2\}$-islands, so at most two pebbles can be moved from each island. Therefore, at most $2(b-j)$ pebbles can be moved from 2-islands to bridges. Hence, at most $2(b-j)+j = 2b-j < 2b$ pebbles can be moved onto bridges, contradicting the fact that $2b$ pebbles were moved onto bridges. ∎

Theorem 4.1 allows us to construct a recursive algorithm for determining the solvability of graphs in $\mathcal{G}_2$. We will show that this algorithm is polynomial for graphs in $\mathcal{G}_{2,k}$ when k is constant.

Consider a configuration C that satisfies the condition of Theorem 4.1. Let r be a vertex of G and let I be an island which is an origin in a minimal r-solution. If I is a 2-island or 3-island, then the vertices of I are used to pebble to some adjacent bridge u, and then never used again. Therefore, instead of considering all possible pebbling sequences from I to u, we can just choose one of them.

If the origin I is a $\{2,2\}$-island, then things are slightly different. If only one pebble is moved from I, the situation is the same as if I were a 2-island. If two pebbles are moved from I, either onto the same adjacent bridge u, or two adjacent bridges u and v, then the vertices of I are never used again, so, as before, we can choose just one such sequence of moves.

Algorithm 4.1 uses these ideas to determine whether every root r in the graph can be pebbled. It maintains a set L of vertices which can be pebbled, which is initialized with every vertex in C_1. It then calls Algorithm 4.2, which recursively tries pebbling moves from every 2-island to every adjacent bridge until every vertex can be pebbled or no moves are possible. At each recursive step, the algorithm checks whether r is adjacent to a 2-island or C' contains an empire, in which case the configuration is solvable. Otherwise, we need to consider configuration C' with fewer pebbles than C.

The following result proves that for every $\{2,2\}$-island I, the recursive call will be made at least once for each vertex in $D_1(I)$ at step 4.2.4 in Algorithm 4.2.

LEMMA 4.2. *Let C be a configuration on a graph $G \in \mathcal{G}_{2,k}$ which contains some vertex r that is not adjacent to any 2-island. If C contains a $\{2,2\}$-island I, then for any vertex $u \in D_1(I)$ there is a vertex $v \in D_1(I)$ such that algorithm* CANDOUBLEPEBBLE(G', I, u, v) *returns* TRUE.

Proof. Let u_1 and u_2 be vertices in I with two or more pebbles and $u \in D_1(I)$. Let

P be a shortest path from $\{u_1, u_2\}$ to u with vertex set $V(P) \subseteq I \cup u$. Without loss of generality, we can assume that $u_1 \in V(P)$. Since P is a shortest path, $u_2 \notin V(P)$. We can move a pebble from u_1 to u using vertices of P and, since $G \in \mathcal{G}_2$, there is a path u_2vr with $v \in D_1(I)$. Therefore, we can move a pebble from u_2 to v and algorithm CanDoublePebble(G', I, u, v) returns true. □

Algorithm 4.1. IsSolvableDiamTwo(G).

```
global Set L
for u ← 0 to n − 1
  do { if u ∈ C_1
       then add u to L                    } 4.1.1
G.UpdateIslands()

return (IsSolvableDiamTwoRec(G))
```

Algorithm 4.2. IsSolvableDiamTwoRec(G).

```
if G.ContainsEmpire()
  then return (true)                      } 4.2.1

for each 2-island I
  do { for each u ∈ D_1(I)
       do {L.Add(u)                       } 4.2.2
if |L| = n
  then return (true)

for each 2-island I
  do { for each u ∈ D_1(I)
       do { G' = G.PebbleFromIsland(I, u)
            if IsSolvableDiamTwoRec(G')   } 4.2.3
              then return (true)

for each {2, 2}-island I
  do { for each u ∈ D_1(I) and v ∈ D_1(I)
       do { if G.CanDoublePebble(G', I, u, v)
            then { if IsSolvableDiamTwoRec(G')   } 4.2.4
                   then return (true)

return (false)
```

The next result provides an upper bound on the time required to determine the solvabilty of a configuration with $|C_2| = l$.

Theorem 4.3. *Let G be a diameter two graph with a configuration C of pebbles such that $|C_2| = l$. Then the solvability of G can be determined in $O(l! \cdot n^{2l-1}m)$ time.*

Proof. We will use induction on $|C_2|$.

Base case. Let C be a configuration with $|C_2| = 1$. Then C is solvable if and only if either every bridge is adjacent to the 2-island or some vertex $v \in C_2$ contains at least four pebbles. Thus, the solvability of C can be determined in $O(n+m)$ time.

Induction step. Let us assume that the theorem is true for $|C_2| \leqslant l$, and let C be a configuration with $|C_2| = l + 1$.

If C contains an empire or every bridge is adjacent to a 2-island, then its solvability can be verified in $O(n+m)$ time. These conditions are checked in steps 4.2.1 and 4.2.2 of Algorithm 4.2.

Otherwise, C does not contain an empire and there is some bridge r that is not adjacent to any 2-island. If there is a minimal r-solution S_r in C, then by Theorem 4.1, there is an island I in G that is an origin for this solution. If I is a 2-island or 3-island, then a subsequence of steps in S_r required to move a pebble to some bridge $u \in D_1(I)$ results in a new configuration C' with $|C'_2| = |C_2| - 1 = l$. If I is a $\{2,2\}$-island that is used in S_r to pebble to bridges u and v (with $u = v$ a possibility), the result of the move(s) from I to u and v is a new configuration C' with $|C'_2| \leqslant |C_2| - 1 = l$. In either case, the resulting graph is solvable by induction.

Since it is unknown which island I is an origin, or which vertex u (or u and v) it pebbles to, we need to consider all of them. For every 2-island and 3-island, we will pebble to each bridge u adjacent to I. There are at most l such islands, and each is adjacent to at most $n - l$ bridges. For each island I and each $u \in D_1(I)$, we need to make the pebbling move and then make a recursive call. This requires $l(n-l)(O(n+m) + T(l-1))$ time, and corresponds to step 4.2.3.

Similarly, for every $\{2,2\}$-island I and pair of bridges u and v (including $u = v$) adjacent to I, we make the pebbling moves required to pebble to both u and v, assuming it is possible, and then make a recursive call. There are at most $l/2$ such islands, each adjacent to at most $(n-l)^2$ pairs of bridges. For each of these, we need to attempt to pebble to both u and v and make a recursive call. This requires $\frac{l}{2}(n-l)^2(O(nm) + T(l-1))$ time, and corresponds to step 4.2.4.

From this, we can see that the complexity of Algorithm 4.1 is

$$T(l) = O(n+m) + l(n-l)(O(n+m) + T(l-1)) + \frac{l}{2}(n-l)^2(O(nm) + T(l-1))$$

$$\leqslant O(l \cdot n^3 m) + l \cdot n^2 T(l-1),$$

where $T(1) = O(n+m)$. It follows that $T(l) = O(l! \cdot n^{2l-1} m)$. ∎

COROLLARY 4.4. *Let $G \in \mathcal{G}_{2,k}$, where k is a constant. Then the solvability of G can be determined in $O(c(k)! \cdot n^{2c(k)-3} m)$ time.*

Proof. Algorithm 4.3 extends Algorithm 4.1 by first determining the connectivity of the graph, and then applying Corollary 2.4. The vertex connectivity of a graph with n vertices and m edges can be determined in $O((n + \min\{k^{5/2}, kn^{3/4}\})m)$ time (see [4]). By Corollary 2.4, the configuration is solvable if $|C_2| \geqslant c(k) = \min\{3k-1, k+4\}$. If $|C_2| < c(k)$, the solvability of the pebbling configuration can be determined by the algorithm IsSolvableDiamTwo in $O(c(k)! \cdot n^{2(c(k)-1)-1} m)$ time. The total time is thus $O(c(k)! \cdot n^{2c(k)-3} m + (n + \min\{k^{5/2}, kn^{3/4}\})m) = O(c(k)! \cdot n^{2c(k)-3} m)$. ∎

Algorithm 4.3. IsSolvableDiamTwoComplete(G).

```
k = G.ComputeConnectivity()
if |C_2| ⩾ min{3k − 1, k + 4}
  then return (TRUE)
  else return (IsSolvableDiamTwo(G))
```

5. Class 1 graphs. The smallest number t, such that any configuration of t pebbles on G is solvable, is called the *pebbling number* of G. Clarke, Hochberg, and Hurlbert [2] provided a complete classification of graphs in $\mathcal{G}_2$. The pebbling number

of any graph in $\mathcal{G}_2$ is either n (these graphs are called *Class* 0) or $n+1$ (called *Class* 1). Figure 5.1 introduces various subsets of $\mathcal{G}_2$ graphs according to class and connectivity. Algorithm 4.3 will determine the solvability of any diameter two graph of Class 0 or 1 in $O(c(k)! \cdot n^{2c(k)-3}m)$ time. For graphs in $\mathcal{D}$ and $\mathcal{F}$, this takes $O(nm)$ and $O(n^7 m)$ time, respectively. In this section, we describe a more efficient technique for determining the solvability of a configuration for graphs of Class 1 using their structural properties. We will show that this approach requires $O(n+m)$ time for graphs in $\mathcal{D}$ and $O(nm)$ time for graphs in $\mathcal{F}$.

	Class	
k	0	1
1	$\emptyset$	$\mathcal{D}$
2	$\mathcal{E}$	$\mathcal{F}$
≥ 3	$\mathcal{H}$	$\emptyset$

FIG. 5.1. *Categorization of diameter two graphs.*

LEMMA 5.1. *Let G have a vertex of degree $n-1$. Then the solvability of G can be determined in $O(n+m)$ time.*

Proof. Let v be a vertex of degree $n-1$. If $C_2 \geqslant 2$, then it is possible to move two pebbles to v, making the configuration solvable. If $C_2 \leqslant 1$, then a configuration is solvable if and only if it contains a vertex with four or more pebbles, every bridge is adjacent to the single 2-island, or every vertex contains one pebble. All of these can be checked in $O(n+m)$ time. ∎

COROLLARY 5.2. *Let $G \in \mathcal{D}$. Then the solvability of G can be determined in $O(n+m)$ time.*

Proof. If $G \in \mathcal{D}$, then G contains a vertex of degree $n-1$. The result follows immediately from Lemma 5.1. ∎

Clarke, Hochberg, and Hurlbert [2] gave a description of the structure of graphs in $\mathcal{F}$, which was corrected by Blasiak and Schmitt [1]. Figure 5.2 shows the structure of all graphs in $\mathcal{F}$. At least two of the edges $p'q'$, $p'r'$, and $q'r'$ must be present. The possibly empty subgraph H_p (similarly for q and r) has all of its vertices adjacent to both q' and r', and each component of H_p has at least one vertex adjacent to p. Finally, each vertex of the subgraph H_c (which also may be empty) must be adjacent to at least two of p', q', and r'. Except for edges within the subgraphs H_p, H_q, H_r, and H_c, no other edges are permitted. Let $H'_p = H_p \cup p$ (similarly for q and r). Note that H'_p is connected if and only if each component of H_p has at least one vertex adjacent to p. Thus we can replace the conditions above with H'_p being nonempty, connected, and each of its vertices being adjacent to both q' and r'.

The existence of a $O(n^5)$ algorithm to determine whether a diameter two graph is Class 1 is implied in [2, 5], but no details are given.

LEMMA 5.3. *Membership in $\mathcal{F}$ can be determined in $O(n^3 m)$ time.*

Proof. We first attempt to identify p', q', and r' by considering all triples of vertices in G. For each triple, we proceed with the following test, quitting if any step fails:

1. Verify that at least two of the edges $p'q'$, $p'r'$, and $q'r'$ are present.
2. Identify the subgraph H'_p by choosing it to be everything that is not in the connected component of $G \setminus \{q', r'\}$ which contains p'. Verify that H'_p is nonempty, connected, and each of its vertices is connected to both q' and r'. Repeat the same process for q and r.

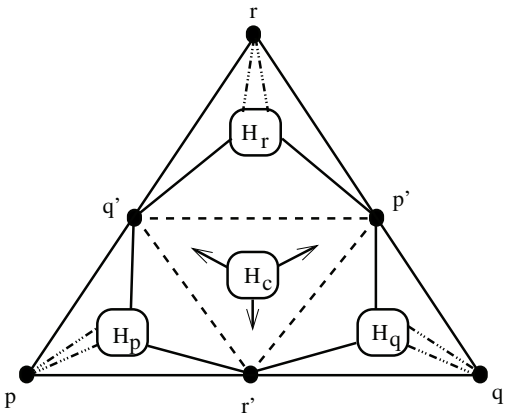


FIG. 5.2. *Family* $\mathcal{F}$.

3. Let H'_c be the set of all vertices not already accounted for. Verify that each one is connected to at least two of p', q', and r'.

Each step takes $O(n+m)$ time. If the test passes for a given choice of p', q', and r', then $G \in \mathcal{F}$. Otherwise, we try the next triple. If this test fails for all triples of vertices in G, $G \notin \mathcal{F}$. Since there are $\binom{n}{3} = O(n^3)$ ways of selecting p', q', and r', the total running time is $O(n^3(n+m)) = O(n^3 m)$. □

By [4], it can be determined in polynomial time whether $G \in \mathcal{G}_{2,1}$ or $G \in \mathcal{G}_{2,2}$ and, by Lemma 5.3, membership in $\mathcal{F}$ can be determined in polynomial time.

COROLLARY 5.4. *Membership in* $\mathcal{D}$, $\mathcal{E}$, $\mathcal{F}$, *and* $\mathcal{H}$ *can be determined in polynomial time.*

The next result describes properties of an unsolvable configuration in $\mathcal{F}$.

LEMMA 5.5. *Let* $G \in \mathcal{F}$. *If* C *is an unsolvable configuration on* G, *then all of the following are true:*

1. *At most one of* p', q', *and* r' *has two or more pebbles.*
2. *At most one vertex in each of* H'_r, H'_p, H'_q, *and* H_c *has two or more pebbles.*
3. *At most two of* H'_r, H'_p, H'_q, *and* H_c *have a vertex with two or more pebbles.*
4. *At most two vertices in* G *have two or more pebbles.*

Proof. Note that any pair of vertices from p', q', and r' forms a cut set, so if statement 1 or 2 is not met, then it is possible to move two pebbles to two of these three vertices. Therefore, by Lemma 2.1, the configuration is solvable. If all three of H_p, H_q, and H_r have a vertex containing two pebbles, it is possible to place two pebbles on any of p', q', and r'. If H_p, H_q, and H_c each have a vertex with two pebbles, then two pebbles can be placed on r', and on either q' or p' (or both). Applying symmetry and Lemma 2.1, the third condition is true. If the first three conditions are satisfied and G contains at least three vertices with two or more pebbles, then one of them must be p', q', or r', and the other two must be from H'_r, H'_p, H'_q, and H'_c with at most one vertex from each. Without loss of generality, assume r' contains two pebbles. Due to symmetry there are four cases to consider:

1. If $u \in H_p$ and $v \in H_q$ each contain two pebbles, four pebbles can be moved to r'.
2. If $u \in H_p$ and $v \in H_r$ each contain two pebbles, two pebbles can be moved to q'.
3. If $u \in H_p$ and $v \in H_c$ each contain two pebbles, either four pebbles can be

moved to r' or two pebbles can be moved to q', depending on which edges among $p'q'$, $p'r'$, and $q'r'$ are present.

4. If $u \in H_r$ and $v \in H_c$ each contain two pebbles, two pebbles can be moved to q' or p', depending on which edges among $p'q'$, $p'r'$, and $q'r'$ are present.

In any of these cases the configuration is solvable. ∎

COROLLARY 5.6. *If $G \in \mathcal{F}$ has a configuration with $|C_2| \geqslant 3$, then C is solvable.*

We need to consider the cases $|C_2| = 1$ and $|C_2| = 2$. The result for $|C_2| = 1$ is obvious, so we state it without proof.

LEMMA 5.7. *Let $G \in \mathcal{G}_2$ and $C_2 = \{u\}$. Then C is solvable if and only if every bridge in C is adjacent to the 2-island or $u \in C_4$.*

We say that a bridge has *potential* s if s pebbles can be moved to it from adjacent islands.

LEMMA 5.8. *Let $G \in \mathcal{G}_2$ and $|C_2| = 2$. Then C is solvable if and only if every bridge is*

1. *adjacent to a 2-island,*
2. *adjacent to a bridge with potential two, or*
3. *adjacent to a desert which is adjacent to a bridge with potential two.*

Proof. The reverse implication is easy to check. For the forward implication, assume C is solvable, and that some bridge r is not adjacent to a 2-island. We need to show that r is adjacent to a bridge with potential two or a desert which is adjacent to a bridge with potential two.

If C contains an empire I, then it is possible to move four pebbles onto some vertex $v \in I$. Since $G \in \mathcal{G}_2$, there is some bridge u adjacent to r and v. Therefore, r is adjacent to a bridge of potential two, and condition 2 is satisfied.

If C does not contain an empire, $|C_2| = 2$ implies that the configuration consists of a $\{2, 2\}$-island, or two 2-islands, either or both of which may be 3-islands. Notice that moving a pebble from a 2-island to a bridge reduces the surplus of the graph by 1, and moving from a 3-island to a bridge reduces the surplus of the graph by 2. In any of these cases, moving pebbles to two different bridges reduces the surplus of the graph to 0, and r cannot be pebbled. Thus, in any r-solution, two pebbles must be moved from the vertices in C_2 onto the same bridge u, reducing the surplus to 1. This implies that r is the only other bridge that can be used in this solution, and u has potential two. Since the configuration is solvable, then r must be adjacent to u or some desert adjacent to u. ∎

The above discussion leads to the following.

THEOREM 5.9. *Let $G \in \mathcal{F}$. Then the solvability of G can be determined in $O(nm)$ time.*

Proof. The conditions of Corollary 5.6 and Lemmas 5.7 and 5.8 can be checked in $O(nm)$ time. ∎

REFERENCES

[1] A. BLASIAK AND J. SCHMITT, *Degree sum conditions in graph pebbling*, Australa. J. Combin., 42 (2008), pp. 83–90.

[2] 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.

[3] T. H. CORMEN, C. E. LEISERSON, R. L. RIVEST, AND C. STEIN, *Introduction to Algorithms*, 2nd ed., MIT Press, Cambridge, MA, 2001.

[4] H. N. GABOW, *Using expander graphs to find vertex connectivity*, J. ACM, 53 (2006), pp. 800–844.

[5] G. HURLBERT, *On Graph Pebbling, Threshold Functions, and Supernormal Posets*, manuscript,

2000.
[6] K. Milans and B. Clark, *The complexity of graph pebbling*, SIAM J. Discrete Math., 20 (2006), pp. 769–798.
[7] Y. Shiloach, *A polynomial solution to the undirected two paths problem*, J. ACM, 27 (1980), pp. 445–456.
[8] N. G. Watson, *The Complexity of Pebbling and Cover Pebbling*, preprint, http://www.arxiv.org/abs/math.CO/0503511, 2005.

SIAM J. DISCRETE MATH.
Vol. 23, No. 2, pp. 647–660

COMPUTING THE DEGREES OF ALL COFACTORS IN MIXED POLYNOMIAL MATRICES*

SATORU IWATA† AND MIZUYO TAKAMATSU‡

Abstract. A mixed polynomial matrix is a polynomial matrix which has two kinds of nonzero coefficients: fixed constants that account for conservation laws and independent parameters that represent physical characteristics. This paper presents an algorithm for computing the degrees of all cofactors simultaneously in a regular mixed polynomial matrix. The algorithm is based on the valuated matroid intersection and all pair shortest paths. The technique is also used for improving the running time of the algorithm for minimizing the index of the differential-algebraic equation in the hybrid analysis for circuit simulation.

Key words. combinatorial matrix theory, degree of cofactor, mixed matrix, polynomial matrix, valuated matroid

AMS subject classifications. 05C50, 15A15, 68Q25

DOI. 10.1137/070706021

1. Introduction.

This paper deals with the computation of the degrees of all cofactors in a polynomial matrix. A polynomial matrix $A(s)$ is said to be *regular* if $A(s)$ is square and $\det A(s)$ is a nonvanishing polynomial. By Cramer's rule, the degrees of cofactors in a regular polynomial matrix determine the degrees of entries of the inverse matrix, which provide useful information for numerical analysis of differential-algebraic equations [2, 3, 14].

A differential-algebraic equation is known to be solvable if it is represented by a regular polynomial matrix whose entries are of degree at most one [1]. For this class of polynomial matrices, Bujakiewicz and van den Bosch [2, 3] proposed an efficient algorithm for finding the degrees of all cofactors under the assumption that coefficients of nonzero entries are independent parameters.

Such a genericity assumption is supported by an argument that the values of physical parameters like resistances in electric circuits are not precise in practice because of noises. However, there do exist exact numbers such as ± 1 that appear in the coefficients of Kirchhoff's conservation laws. This observation led Murota and Iri [15] to introduce the notion of a *mixed matrix*, which is a constant matrix that consists of two kinds of numbers as follows.

Accurate numbers (fixed constants): Numbers that account for conservation laws are precise in values. These numbers should be treated numerically.

Inaccurate numbers (independent parameters): Numbers that represent physical characteristics are not precise in values. These numbers should be treated combinatorially as nonzero parameters without reference to their nominal values. Since each such nonzero entry often comes from a single physical device, the parameters are assumed to be independent.

*Received by the editors October 22, 2007; accepted for publication (in revised form) May 26, 2008; published electronically March 4, 2009. This work is supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan.
http://www.siam.org/journals/sidma/23-2/70602.html
†Research Institute for Mathematical Sciences, Kyoto University, Kyoto 606-8502, Japan (iwata@kurims.kyoto-u.ac.jp).
‡Graduate School of Information Science and Technology, University of Tokyo, Tokyo 113-8656, Japan (mizuyo_takamatsu@mist.i.u-tokyo.ac.jp).

In order to deal with dynamical systems, it is natural to consider the polynomial matrix version, which is called a *mixed polynomial matrix* [14].

For a mixed polynomial matrix $A(s)$, Murota [13] devised an algorithm for computing the maximum degree of minors of order r:

$$\delta_r(A) = \max_{I,J}\{\deg \det A[I,J] \mid |I| = |J| = r\},$$

where $A[I,J]$ denotes the submatrix of $A(s)$ indexed by row set I and column set J, as an application of the valuated matroid intersection [11, 12]. This algorithm is also used for computing $\delta_r(\bar{A})$ for a polynomial matrix $\bar{A}(s)$ that is obtained by substituting specific numerical values to independent parameters of a mixed polynomial matrix [8].

For a regular mixed polynomial matrix, Murota's algorithm provides the degree of the determinant and the maximum degree of cofactors. A straightforward approach to the degrees of all cofactors in an $n \times n$ mixed polynomial matrix requires n^2 applications of this algorithm, which results in high order time complexity.

In this paper, we present an efficient algorithm for finding the degrees of all cofactors in a regular mixed polynomial matrix simultaneously, which is an extension of the result of Bujakiewicz and van den Bosch [2, 3]. The proposed algorithm first applies Murota's algorithm for the degree of the determinant. Then, from the obtained solution, it finds the optimal value of the associated valuated matroid intersection, which coincides with the degree of each cofactor. This can be done at once by using all pair shortest paths algorithm.

The time complexity is the same as that of the algorithm for the degree of the determinant described in [13], because it is dominated by the computation in the first step. The technique is also used to improve the complexity of the algorithm in [9] for finding an optimal hybrid analysis in which the index of the differential-algebraic equation to be solved attains the minimum.

The organization of this paper is as follows. Section 2 provides preliminaries on mixed polynomial matrices and valuated matroids. In section 3, we describe the algorithm of Murota for computing the degree of the determinant of a regular mixed polynomial matrix. Section 4 gives a characterization of the degree of a cofactor. We present an algorithm for computing the degrees of all cofactors simultaneously in a regular mixed polynomial matrix and analyze its running time in section 5. Finally, in section 6, we discuss a similar problem which appears in the index minimization of the differential-algebraic equation in the hybrid analysis.

2. Preliminaries. This section is devoted to preliminaries on mixed polynomial matrices and valuated matroids. Valuated matroids are combinatorial abstractions of polynomial matrices.

For a polynomial $a(s)$, we denote the degree of $a(s)$ by $\deg a$, where $\deg 0 = -\infty$ by convention. For a polynomial matrix $A(s)$, we denote by $A[I,J]$ the submatrix of $A(s)$ with row set $I \subseteq R$ and column set $J \subseteq C$, where R and C are the row set and the column set of $A(s)$, respectively. The (i,j) entry of $A(s)$ is denoted by $A_{ij}(s)$.

A *generic matrix* is a matrix in which each nonzero entry is an independent parameter. A matrix $A(s)$ is called a *mixed polynomial matrix* if $A(s)$ is given by $A(s) = Q(s) + T(s)$ with a pair of polynomial matrices $Q(s) = \sum_{h=0}^{N} s^h Q_h$ and $T(s) = \sum_{h=0}^{N} s^h T_h$ that satisfy the following two conditions.

(MP-Q) The coefficients Q_h $(h = 0, \ldots, N)$ in $Q(s)$ are constant matrices.

(MP-T) The coefficients T_h $(h = 0, \ldots, N)$ in $T(s)$ are generic matrices.

A *layered mixed polynomial matrix* (or an *LM-polynomial matrix* for short) is defined to be a mixed polynomial matrix such that $Q(s)$ and $T(s)$ satisfying (MP-Q) and (MP-T) have disjoint nonzero rows. An LM-polynomial matrix $A(s)$ is expressed by $A(s) = \binom{Q(s)}{T(s)}$.

Dress and Wenzel [4] defined a *valuated matroid* to be a triple $\mathbf{M} = (V, \mathcal{B}, \omega)$ of a finite set V, a nonempty family $\mathcal{B} \subseteq 2^V$, and a function $\omega : \mathcal{B} \to \mathbf{R}$ that satisfy the following axiom (VM).

(VM) For any $B, B' \in \mathcal{B}$ and $u \in B \setminus B'$, there exists $v \in B' \setminus B$ such that $B \setminus \{u\} \cup \{v\} \in \mathcal{B}$, $B' \cup \{u\} \setminus \{v\} \in \mathcal{B}$, and $\omega(B) + \omega(B') \leq \omega(B \setminus \{u\} \cup \{v\}) + \omega(B' \cup \{u\} \setminus \{v\})$.

The function ω is called a *valuation*. For $B \in \mathcal{B}, u \in B$, and $v \in V \setminus B$, we define

$$\omega(B, u, v) = \omega(B \setminus \{u\} \cup \{v\}) - \omega(B).$$

By convention, we put $\omega(B, u, v) = -\infty$ if $B \setminus \{u\} \cup \{v\} \notin \mathcal{B}$.

The local optimality for the valuation implies the global optimality as follows.

THEOREM 2.1 (see [14, Theorem 5.2.7]). *A base $B \in \mathcal{B}$ satisfies $\omega(B) \geq \omega(B')$ for any $B' \in \mathcal{B}$ if and only if $\omega(B, u, v) \leq 0$ holds for any $u \in B$ and $v \in V \setminus B$.*

For $B \in \mathcal{B}$ and $B' \subseteq V$, we consider a bipartite graph, called the *exchangeability graph*, $G(B, B') = (B \setminus B', B' \setminus B; H)$ with

$$H = \{(u, v) \mid u \in B \setminus B', v \in B' \setminus B, B \setminus \{u\} \cup \{v\} \in \mathcal{B}\}.$$

We denote by $\hat{\omega}(B, B')$ the maximum weight of a perfect matching in $G(B, B')$, with respect to the edge weight $\omega(B, u, v)$, i.e.,

$$\hat{\omega}(B, B') = \max \left\{ \sum_{(u,v) \in M} \omega(B, u, v) \mid M \text{ is a perfect matching in } G(B, B') \right\}.$$

A necessary and sufficient condition for the unique existence of the maximum-weight perfect matching in $G(B, B')$ is given as follows.

LEMMA 2.2 (see [14, Lemma 5.2.32]). *Let $B \in \mathcal{B}$ and $B' \subseteq V$ with $|B' \setminus B| = |B \setminus B'| = h$. There exists exactly one maximum-weight perfect matching in $G(B, B')$ if and only if there exist $q : (B \setminus B') \cup (B' \setminus B) \to \mathbf{R}$ and indexings of elements of $B \setminus B'$ and $B' \setminus B$, say $B \setminus B' = \{u_1, \ldots, u_h\}$ and $B' \setminus B = \{v_1, \ldots, v_h\}$, such that*

$$\omega(B, u_j, v_i) + q(u_j) - q(v_i) \begin{cases} = 0 & (1 \leq i = j \leq h), \\ \leq 0 & (1 \leq i < j \leq h), \\ < 0 & (1 \leq j < i \leq h), \end{cases} \tag{2.1}$$

where the latter condition implies that $\hat{\omega}(B, B') = \sum_{i=1}^{h} q(v_i) - \sum_{i=1}^{h} q(u_i)$.

The following lemma is called the "unique-max lemma."

LEMMA 2.3 (see [14, Lemma 5.2.35]). *Let $B \in \mathcal{B}$ and $B' \subseteq V$ with $|B'| = |B|$. If there exists exactly one maximum-weight perfect matching in $G(B, B')$, then $B' \in \mathcal{B}$ and $\omega(B') = \omega(B) + \hat{\omega}(B, B')$.*

Murota [11] introduced the *valuated independent assignment problem* as a generalization of the independent assignment problem [7]. The valuated independent assignment problem VIAP(r) parametrized by an integer r is as follows [14, p. 307].

[VIAP(r)] Given a bipartite graph $G = (V^+, V^-; E)$ with vertex sets V^+, V^- and edge set E, a pair of valuated matroids $\mathbf{M}^+ = (V^+, \mathcal{B}^+, \omega^+)$ and $\mathbf{M}^- = (V^-, \mathcal{B}^-, \omega^-)$, and a weight function $w : E \to \mathbf{R}$, find a triple (M, B^+, B^-) that maximizes

$$\Omega(M, B^+, B^-) := w(M) + \omega^+(B^+) + \omega^-(B^-),$$

where $w(M) = \sum\{w(a) \mid a \in M\}$, subject to the constraint that $M \subseteq E$ is a matching of size r and

$$\partial^+ M \subseteq B^+ \in \mathcal{B}^+, \quad \partial^- M \subseteq B^- \in \mathcal{B}^-, \tag{2.2}$$

where $\partial^+ M$ and $\partial^- M$ denote the set of vertices in V^+ and V^- incident to M, respectively.

An augmenting path algorithm for solving VIAP(r) has been developed in [12], where the unique-max lemma plays a key role.

3. Degree of determinant. Let $\tilde{A}(s) = \tilde{Q}(s) + \tilde{T}(s)$ be an $n \times n$ regular mixed polynomial matrix with row set $\tilde{R}$ and column set $\tilde{C}$. In this section, we expound that the computation of

$$\delta_r(\tilde{A}) = \max_{I,J}\{\deg \det \tilde{A}[I, J] \mid |I| = |J| = r\},$$

the highest degree of a minor of order r, is reduced to solving VIAP(r) [13, 14].

Let us define

$$g_i = \max_{j \in \tilde{C}} \deg \tilde{Q}_{ij}(s) \quad (i \in \tilde{R}). \tag{3.1}$$

We now construct an associated $2n \times 2n$ LM-polynomial matrix:

$$A(s) = \begin{pmatrix} Q(s) \\ T(s) \end{pmatrix} = \begin{array}{c} \\ R_Q \\ R_T \end{array} \begin{array}{c} \begin{array}{cc} \tilde{R} & \tilde{C} \end{array} \\ \begin{pmatrix} D_Q(s) & \tilde{Q}(s) \\ -D_T(s) & \tilde{T}(s) \end{pmatrix} \end{array} \tag{3.2}$$

with column set $C = \tilde{R} \cup \tilde{C}$ and row set $R = R_Q \cup R_T$, where R_Q and R_T are disjoint copies of $\tilde{R}$. For each $i \in \tilde{R}$, we denote its copies by $i_Q \in R_Q$ and $i_T \in R_T$. Both $D_Q(s)$ and $D_T(s)$ are diagonal matrices. For each $i \in \tilde{R}$, the (i_Q, i) entry of $D_Q(s)$ is s^{g_i}, and the (i_T, i) entry of $D_T(s)$ is $t_i s^{g_i}$, where t_i is a new independent parameter.

For an LM-polynomial matrix $A(s) = \binom{Q(s)}{T(s)}$ in general, let R_Q and R_T denote the row sets of $Q(s)$ and $T(s)$. We also denote $|R_Q|$ and $|R_T|$ by m_Q and m_T, respectively. The degree of $\det A$ is expressed as follows.

THEOREM 3.1 (see [14, Theorem 6.2.5]). *For a regular LM-polynomial matrix $A(s) = \binom{Q(s)}{T(s)}$, we have*

$$\deg \det A = \max_{J \subseteq C, |J| = m_Q} \{\deg \det Q[R_Q, J] + \deg \det T[R_T, C \setminus J]\}.$$

The degrees of $\det Q[R_Q, J]$ and $\det T[R_T, C \setminus J]$ correspond to the valuation and the maximum weight of bipartite matchings, respectively. For $r = 0, 1, \ldots, m_T$, we define

$$\delta_r^{\mathrm{LM}}(A) = \max_{I,J}\{\deg \det A[R_Q \cup I, J] \mid I \subseteq R_T, J \subseteq C, |I| = r, |J| = m_Q + r\},$$

which designates the highest degree of a minor of order m_Q+r with row set containing R_Q. Note that we have $\delta^{\mathrm{LM}}_{m_T}(A) = \deg\det A$ for a square LM-polynomial matrix $A(s)$.

For an associated LM-polynomial matrix $A(s)$ with an $n \times n$ mixed polynomial matrix $\tilde{A}(s)$, we have $m_Q = m_T = n$. The relation between $\delta_r(\tilde{A})$ and $\delta^{\mathrm{LM}}_r(A)$ is as follows.

LEMMA 3.2 (see [14, Lemma 6.2.6]). *Let $\tilde{A}(s) = \tilde{Q}(s) + \tilde{T}(s)$ be an $n \times n$ mixed polynomial matrix with row set $\tilde{R}$. We denote by $A(s)$ the associated LM-polynomial matrix defined by* (3.1) *and* (3.2). *For an integer r with $0 \le r \le n$, we have*

$$\delta_r(\tilde{A}) = \delta^{\mathrm{LM}}_r(A) - \sum_{i\in\tilde{R}} g_i. \tag{3.3}$$

Remark 3.3. In fact, (3.3) holds for an associated LM-polynomial matrix defined by (3.2) if each g_i satisfies $g_i \ge \max_{j\in\tilde{C}} \deg \tilde{Q}_{ij}(s)$.

Example 3.4. Consider a mixed polynomial matrix

$$\tilde{A} = \begin{pmatrix} 1 & 0 & s \\ 0 & 1 & 0 \\ 0 & t_1 s & 1+t_2 s \end{pmatrix} = \begin{pmatrix} 1 & 0 & s \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & t_1 s & t_2 s \end{pmatrix}$$

with row set $\tilde{R} = \{x_1, x_2, x_3\}$ and column set $\tilde{C} = \{y_1, y_2, y_3\}$. The associated LM-polynomial matrix defined by (3.1) and (3.2) is

$$A = \begin{array}{c} \\ x_{1Q} \\ x_{2Q} \\ x_{3Q} \\ x_{1T} \\ x_{2T} \\ x_{3T} \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & y_1 & y_2 & y_3 \end{array} \\ \begin{pmatrix} s & 0 & 0 & 1 & 0 & s \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ -t_3 s & 0 & 0 & 0 & 0 & 0 \\ 0 & -t_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & -t_5 & 0 & t_1 s & t_2 s \end{pmatrix} \end{array}.$$

Then we have $\delta_3(\tilde{A}) = 1$ and $\delta^{\mathrm{LM}}_3(A) = 2$, which satisfy (3.3).

By Lemma 3.2, $\delta_r(\tilde{A})$ is determined from $\delta^{\mathrm{LM}}_r(A)$. We now describe how to reduce the computation of $\delta^{\mathrm{LM}}_r(A)$ to VIAP(r).

Let $\mathbf{M}_Q = (C, \mathcal{B}_Q, \omega_Q)$ be a valuated matroid defined by

$$\mathcal{B}_Q = \{B \subseteq C \mid \det Q[R_Q, B] \neq 0\}, \quad \omega_Q(B) = \deg\det Q[R_Q, B] \quad (B \in \mathcal{B}_Q).$$

Consider a bipartite graph $G = (V^+, V^-; E)$ with $V^+ = R_T$, $V^- = C$, and $E = \{(i,j) \mid i \in R_T, j \in C, T_{ij}(s) \neq 0\}$. Let VIAP($A; r$) denote VIAP($r$) defined on G as follows. The valuated matroids $\mathbf{M}^+ = (V^+, \mathcal{B}^+, \omega^+)$ and $\mathbf{M}^- = (V^-, \mathcal{B}^-, \omega^-)$ attached to V^+ and V^- are defined by

$$\mathcal{B}^+ = \{R_T\}, \quad \omega^+(R_T) = 0$$

and

$$\mathcal{B}^- = \{B \subseteq C \mid C \setminus B \in \mathcal{B}_Q\}, \quad \omega^-(B) = \omega_Q(C \setminus B) \quad (B \in \mathcal{B}^-).$$

The weight $w(a)$ of an arc $a = (i,j) \in E$ is given by $w(a) = \deg T_{ij}(s)$. Figure 3.1 illustrates G of Example 3.4.

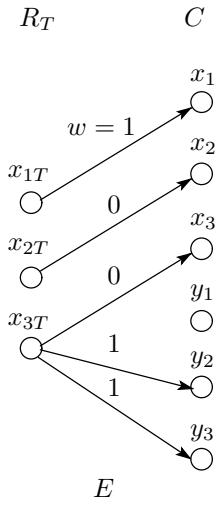


FIG. 3.1. *A bipartite graph G of Example* 3.4.

A pair (M, B) of a matching $M \subseteq E$ and a base $B \in \mathcal{B}^-$ is called *feasible* for VIAP$(A; r)$ if $|M| = r$ and $\partial^- M \subseteq B$. The value of a feasible pair (M, B) is given by

$$\begin{aligned}\Omega_r(M, B) &= w(M) + \omega^+(R_T) + \omega^-(B)\\ &= w(M) + \omega_Q(C \setminus B)\\ &= \deg \det Q[R_Q, C \setminus B] + \sum_{(i,j)\in M} \deg T_{ij}(s).\end{aligned}$$

A feasible pair that maximizes $\Omega_r(M, B)$ is called *optimal* for VIAP$(A; r)$. The following theorem shows that the optimal value of VIAP$(A; r)$ coincides with $\delta_r^{\mathrm{LM}}(A)$.

THEOREM 3.5 (see [14, Theorem 6.2.8]). *For a square LM-polynomial matrix $A(s)$ and an integer r with $0 \le r \le m_T$, we have*

$$\delta_r^{\mathrm{LM}}(A) = \max\{\Omega_r(M, B) \mid (M, B) \text{ is feasible for VIAP}(A; r)\},$$

where the right-hand side is defined to be $-\infty$ if there exists no feasible pair (M, B).

We now describe the algorithm for computing $\delta_r^{\mathrm{LM}}(A)$, proposed by Murota [13, 14]. The algorithm solves VIAP$(A; r)$ successively for $r = 0, 1, \ldots, m_T$. It maintains a feasible pair (M, B) that maximizes $\Omega_r(M, B)$. Note that this algorithm works even if $A(s)$ is not regular.

Let us denote the reorientation of $a \in E$ by a°. With reference to G and (M, B), we construct an auxiliary graph $G^* = (R_T \cup C, E^*)$ with arc set $E^* = E \cup E^- \cup M^\circ$, where

$$E^- = \{(v, u) \mid u \in B, v \in C \setminus B, B \setminus \{u\} \cup \{v\} \in \mathcal{B}^-\}, \quad M^\circ = \{a^\circ \mid a \in M\}.$$

Note that the arcs in E^- have both ends in C and that the arcs in M° are directed from C to R_T. The arc length $\gamma : E^* \to \mathbf{Z}$ is defined by

$$\gamma(a) = \begin{cases} -w(a) & (a \in E),\\ w(a^\circ) & (a \in M^\circ),\\ -\omega^-(B, u, v) & (a = (v, u) \in E^-), \end{cases} \tag{3.4}$$

where $\omega^-(B,u,v) = \omega^-(B \setminus \{u\} \cup \{v\}) - \omega^-(B)$. We put $S^+ = R_T \setminus \partial^+ M$ and $S^- = B \setminus \partial^- M$. Let $\partial^+ a$ and $\partial^- a$ denote the initial and terminal vertices of a, respectively. Then the following fact holds.

THEOREM 3.6 (see [14, Theorem 5.2.62]). *Let (M,B) be an optimal pair for* VIAP$(A;r)$ *and P be a shortest path from S^+ to S^- with respect to the arc length γ in G^* having the smallest number of arcs. Then $(\hat{M},\hat{B})$ defined by*

$$\hat{M} = M \setminus \{a \in M \mid a^\circ \in P \cap M^\circ\} \cup (P \cap E), \tag{3.5}$$

$$\hat{B} = B \setminus \{\partial^- a \mid a \in P \cap E^-\} \cup \{\partial^+ a \mid a \in P \cap E^-\} \tag{3.6}$$

is optimal for VIAP$(A;r+1)$.

Theorem 3.6 leads to the following algorithm for computing the degree of the determinant of a regular LM-polynomial matrix.

ALGORITHM FOR DEGREE OF DETERMINANT.

Step 1: Find a maximum-weight base $B \in \mathcal{B}^-$ with respect to ω^-. Put $M := \emptyset$.

Step 2: Repeat (2-1)–(2-3) until $|M| = m_T$.

(2-1) Construct an auxiliary graph G^* with respect to (M,B).

(2-2) Find a shortest path P having the smallest number of arcs from S^+ to S^- with respect to the arc length γ in G^*.

(2-3) Update (M,B) according to (3.5) and (3.6).

At each stage of this algorithm, it holds that $\delta_r^{\mathrm{LM}}(A) = \Omega_r(M,B)$ for $r = |M|$. At the end of the algorithm, we obtain an optimal pair (M,B) for VIAP$(A;n)$.

4. Degree of cofactor. Let $\tilde{A}(s)$ be an $n \times n$ regular mixed polynomial matrix and $A(s)$ be the associated LM-polynomial matrix defined by (3.1) and (3.2). In this section, we discuss the degree of a cofactor in $\tilde{A}(s)$. We first show that the degree of a cofactor in $\tilde{A}(s)$ is determined by that of the corresponding cofactor in $A(s)$.

LEMMA 4.1. *Let $\tilde{A}(s)$ be an $n \times n$ mixed polynomial matrix and $A(s)$ be the associated LM-polynomial matrix defined by* (3.1) *and* (3.2). *For $k \in \tilde{R}$ and $l \in \tilde{C}$, we have*

$$\deg\det \tilde{A}[\tilde{R} \setminus \{k\}, \tilde{C} \setminus \{l\}] = \deg\det A[R \setminus \{k_T\}, C \setminus \{l\}] - \sum_{i \in \tilde{R}} g_i. \tag{4.1}$$

Proof. Applying Remark 3.3 to a mixed polynomial matrix $\tilde{A}[\tilde{R} \setminus \{k\}, \tilde{C} \setminus \{l\}]$ and an LM-polynomial matrix $A[R \setminus \{k_Q, k_T\}, C \setminus \{k,l\}]$, we have

$$\deg\det \tilde{A}[\tilde{R} \setminus \{k\}, \tilde{C} \setminus \{l\}] = \deg\det A[R \setminus \{k_Q, k_T\}, C \setminus \{k,l\}] - \sum_{i \in \tilde{R} \setminus \{k\}} g_i.$$

Since the degree of the (k_Q, k) entry of A is g_k and $A[R \setminus \{k_Q, k_T\}, \{k\}] = O$, it follows that

$$\deg\det A[R \setminus \{k_Q, k_T\}, C \setminus \{k,l\}] = \deg\det A[R \setminus \{k_T\}, C \setminus \{l\}] - g_k.$$

Thus we obtain (4.1). ∎

By Lemma 4.1, it suffices to compute $\deg\det A[R \setminus \{k_T\}, C \setminus \{l\}]$ for $k \in \tilde{R}$ and $l \in \tilde{C}$. We now define the following problem.

[DOC$(A; k_T, l)$] Find a pair (M,B) of a matching $M \subseteq E$ and a base $B \in \mathcal{B}^-$ maximizing $w(M) + \omega^-(B)$ subject to

$$\partial^+ M = R_T \setminus \{k_T\}, \quad \partial^- M = B \setminus \{l\}, \quad l \in B. \tag{4.2}$$

A pair (M, B) that satisfies (4.2) is *feasible* for $\mathrm{DOC}(A; k_T, l)$. Similarly to Theorem 3.5, the degree of $\det A[R \setminus \{k_T\}, C \setminus \{l\}]$ coincides with the optimal value of $\mathrm{DOC}(A; k_T, l)$. The following proposition gives a sufficient condition for the optimality of $\mathrm{DOC}(A; k_T, l)$.

PROPOSITION 4.2. *A feasible pair (M, B) for $\mathrm{DOC}(A; k_T, l)$ is optimal if there exists a pair of vectors $p : R_T \to \mathbf{R}$ and $q : C \to \mathbf{R}$ with $q(l) = 0$ such that*

(i) $w(a) - p(\partial^+ a) + q(\partial^- a) \leq 0$ *holds for* $a \in E$,

(ii) $w(a) - p(\partial^+ a) + q(\partial^- a) = 0$ *holds for* $a \in M$,

(iii) B *maximizes* $\omega^-[-q]$, *where* $\omega^-[-q](B) \equiv \omega^-(B) - \sum_{u \in B} q(u)$.

Proof. For any feasible pair (M', B') for $\mathrm{DOC}(A; k_T, l)$, we show that

$$w(M') + \omega^-(B') \leq w(M) + \omega^-(B). \tag{4.3}$$

By (i) and the feasibility of (M', B'), we have

$$w(M') + \omega^-(B') \leq p(\partial^+ M') - q(\partial^- M') + \omega^-(B') = p(R_T \setminus \{k_T\}) - q(B' \setminus \{l\}) + \omega^-(B'),$$

where $p(I) = \sum_{i \in I} p(i)$ and $q(J) = \sum_{j \in J} q(j)$. It follows from $q(l) = 0$ that

$$-q(B' \setminus \{l\}) + \omega^-(B') = -q(B') + \omega^-(B') = \omega^-[-q](B').$$

By (iii), we have $\omega^-[-q](B') \leq \omega^-[-q](B)$. Thus we obtain

$$w(M') + \omega^-(B') \leq p(R_T \setminus \{k_T\}) + \omega^-[-q](B) = p(R_T \setminus \{k_T\}) + \omega^-(B) - q(B),$$

which implies (4.3) by (ii) and $q(l) = 0$. □

With reference to an optimal pair (M, B) for $\mathrm{VIAP}(A; n)$, we construct the auxiliary graph G^*. For each pair of vertices u and v, let $d(u, v)$ denote the shortest path distance from u to v with respect to the arc length γ in G^*. If there exists no path from u to v, then we put $d(u, v) = \infty$. The degree of a cofactor is now characterized as follows.

THEOREM 4.3. *Let (M, B) be an optimal pair for $\mathrm{VIAP}(A; n)$. Then we have*

$$\deg \det A[R \setminus \{k_T\}, C \setminus \{l\}] = \Omega_n(M, B) - d(l, k_T)$$

for any $k_T \in R_T$ and $l \in C$.

Let (M, B) be an optimal pair for $\mathrm{VIAP}(A; n)$ and P be a shortest path from l to k_T with respect to the arc length γ in G^* having the smallest number of arcs. We update (M, B) to $(\hat{M}, \hat{B})$ according to (3.5) and (3.6). Let $\{(v_i, u_i) \mid i = 1, \ldots, h\} = P \cap E^-$, where $h = |P \cap E^-|$, and the indices are chosen so that $v_h, u_h, \ldots, v_1, u_1$ appear on P in this order. In order to prove Theorem 4.3, we make use of the following lemma.

LEMMA 4.4. *Let $G(B, \hat{B})$ be the exchangeability graph with respect to the valuated matroid $(V^-, \mathcal{B}^-, \omega^-)$. Then there exists exactly one maximum-weight perfect matching in $G(B, \hat{B})$. Moreover, we have*

$$\hat{\omega}^-(B, \hat{B}) = \sum_{i=1}^{h} d(l, v_i) - \sum_{i=1}^{h} d(l, u_i). \tag{4.4}$$

Proof. Consider $q(v) = d(l, v)$ for each $v \in V^-$. Then we have $q(v_i) - \omega^-(B, u_j, v_i) \geq q(u_j)$ for any $(v_i, u_j) \in E^-$. The equality holds if $i = j$ and the strict inequality

does if $j < i$. Hence, by Lemma 2.2, there exists exactly one maximum-weight perfect matching in $G(B, \hat{B})$, and (4.4) holds. □

We are now ready to complete the proof of Theorem 4.3. Note that $(\hat{M}, \hat{B})$ is feasible for $\mathrm{DOC}(A; k_T, l)$. We claim that $(\hat{M}, \hat{B})$ is optimal for $\mathrm{DOC}(A; k_T, l)$.

Consider $p(u) = d(l, u)$ for $u \in V^+$ and $q(v) = d(l, v)$ for $v \in V^-$. We show that p, q, and $(\hat{M}, \hat{B})$ satisfy the following:

(i) $w(a) - p(\partial^+ a) + q(\partial^- a) \leq 0$ holds for $a \in E$,

(ii) $w(a) - p(\partial^+ a) + q(\partial^- a) = 0$ holds for $a \in \hat{M}$,

(iii) $\hat{B}$ maximizes $\omega^-[-q]$,

which is the condition in Proposition 4.2 applied to $(\hat{M}, \hat{B})$. The definition of p and q implies that (i) and (ii) hold. By Lemmas 2.3 and 4.4, we have

$$\omega^-(\hat{B}) = \omega^-(B) + \hat{\omega}^-(B, \hat{B}). \tag{4.5}$$

It follows from (4.4) that

$$\hat{\omega}^-(B, \hat{B}) = \sum_{v \in \hat{B} \setminus B} q(v) - \sum_{u \in B \setminus \hat{B}} q(u) = q(\hat{B} \setminus B) - q(B \setminus \hat{B}) = q(\hat{B}) - q(B).$$

Thus we obtain $\omega^-(\hat{B}) - q(\hat{B}) = \omega^-(B) - q(B)$. This can be written as $\omega^-[-q](\hat{B}) = \omega^-[-q](B)$. By the definition of q, for any $u \in B$ and $v \in V^- \setminus B$, we have $q(v) - \omega^-(B, u, v) \geq q(u)$, which implies that $\omega^-(B) \geq \omega^-(B \setminus \{u\} \cup \{v\}) + q(u) - q(v)$. Hence

$$\begin{aligned} \omega^-[-q](B \setminus \{u\} \cup \{v\}) &= \omega^-(B \setminus \{u\} \cup \{v\}) - q(B) + q(u) - q(v) \\ &\leq \omega^-(B) - q(B) = \omega^-[-q](B) \end{aligned}$$

holds. Since the triple $(V^-, \mathcal{B}^-, \omega^-[-q])$ is a valuated matroid, it follows from Theorem 2.1 that $\omega^-[-q](B') \leq \omega^-[-q](B) = \omega^-[-q](\hat{B})$ holds for any $B' \in \mathcal{B}^-$, which implies (iii). Therefore, by Proposition 4.2, $(\hat{M}, \hat{B})$ is optimal for $\mathrm{DOC}(A; k_T, l)$.

Since the degree of $\det A[R \setminus \{k_T\}, C \setminus \{l\}]$ coincides with the optimal value of $\mathrm{DOC}(A; k_T, l)$, we have $\deg \det A[R \setminus \{k_T\}, C \setminus \{l\}] = w(\hat{M}) + \omega^-(\hat{B})$. It follows from (3.4) and (3.5) that

$$w(\hat{M}) = w(M) - \sum_{a^\circ \in P \cap M^\circ} w(a) + \sum_{a \in P \cap E} w(a) = w(M) - \sum_{a \in P \cap M^\circ} \gamma(a) - \sum_{a \in P \cap E} \gamma(a).$$

By (4.4) and (4.5), we obtain

$$\omega^-(\hat{B}) = \omega^-(B) + \hat{\omega}^-(B, \hat{B}) = \omega^-(B) - \sum_{a \in P \cap E^-} \gamma(a).$$

Therefore, we have $w(\hat{M}) + \omega^-(\hat{B}) = w(M) + \omega^-(B) - \sum_{a \in P} \gamma(a) = \Omega_n(M, B) - d(l, k_T)$. Thus $\deg \det A[R \setminus \{k_T\}, C \setminus \{l\}] = \Omega_n(M, B) - d(l, k_T)$ holds, which completes the proof of Theorem 4.3.

Example 4.5. For the LM-polynomial matrix of Example 3.4, Figure 4.1 exhibits an optimal pair (M, B) for $\mathrm{VIAP}(A; 3)$ and an auxiliary graph G^* with

$$M = \{(x_{1T}, x_1), (x_{2T}, x_2), (x_{3T}, y_3)\} \quad \text{and} \quad B = \{x_1, x_2, y_3\}.$$

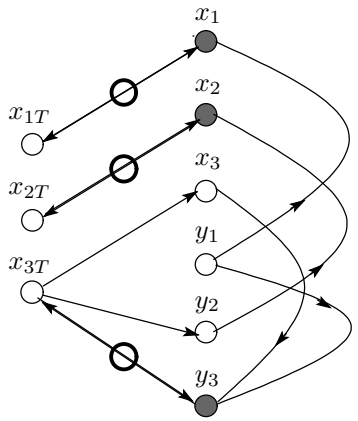


FIG. 4.1. *An auxiliary graph G^* of Example* 3.4, *where $\ominus$ and $\bullet$ denote arcs in M and vertices in B, respectively.*

Then we have $\Omega_3(M, B) = 2$. Consider the degree of $\det A[R \setminus \{x_{2T}\}, C \setminus \{y_1\}]$. A shortest path P from y_1 to x_{2T} in G^* is

$$P = \{(y_1, y_3), (y_3, x_{3T}), (x_{3T}, y_2), (y_2, x_2), (x_2, x_{2T})\}$$

and its shortest path distance is $d(y_1, x_{2T}) = \gamma(P) = -1$. It follows from Theorem 4.3 that

$$\deg \det A[R \setminus \{x_{2T}\}, C \setminus \{y_1\}] = \Omega_3(M, B) - d(y_1, x_{2T}) = 3.$$

5. Degrees of all cofactors. In this section, we present an algorithm for computing the degrees of all cofactors simultaneously and analyze its running time.

Theorem 4.3 suggests the following algorithm for computing the degrees of all cofactors in an $n \times n$ regular mixed polynomial matrix $\tilde{A}(s) = \tilde{Q}(s) + \tilde{T}(s)$. The output of this algorithm is a matrix Ψ whose (k, l) entry, denoted by ψ_{kl}, is the degree of the cofactor $\det \tilde{A}[\tilde{R} \setminus \{k\}, \tilde{C} \setminus \{l\}]$.

ALGORITHM FOR DEGREES OF ALL COFACTORS.

Step 1: Construct the $2n \times 2n$ associated LM-polynomial matrix $A(s)$ defined by (3.1) and (3.2).

Step 2: Find an optimal pair (M, B) for VIAP$(A; n)$ by the algorithm for degree of determinant. Construct an auxiliary graph G^* with respect to (M, B).

Step 3: Compute the shortest path distances for all pairs of $k_T \in R_T$ and $l \in \tilde{C}$. For each $k \in \tilde{R}$ and $l \in \tilde{C}$, set $\psi_{kl} := \Omega_n(M, B) - d(l, k_T) - \sum_{i \in \tilde{R}} g_i$.

Step 4: Return Ψ.

We now discuss the running time of the algorithm for degrees of all cofactors. In Step 3, we can compute the shortest path distances for all pairs by the *Warshall–Floyd method* [5, 16] in $O(n^3)$ time. This is dominated by the algorithm for degree of determinant in Step 2. Thus the overall time complexity of the algorithm for degrees of all cofactors is the same as that of the algorithm for degree of determinant.

In order to reflect the dimensional consistency in conservation laws, Murota [10] introduced the following assumption.

(MP-Q2) Every nonvanishing minor of $\tilde{Q}(s)$ is a monomial in s.

For example, consider a linear time-invariant electric circuit. As for the coefficient matrix $\tilde{A}(s)$ of circuit equations, which consist of Kirchhoff's conservation laws (KCL

and KVL) and constitutive equations, we assume that the physical parameters are independent. Then, $\tilde{A}(s)$ is an LM-polynomial matrix that satisfies (MP-Q2).

The assumption (MP-Q2) holds if and only if

$$\tilde{Q}(s) = D_R(s)\tilde{Q}(1)D_C(s) \tag{5.1}$$

for some diagonal matrices $D_R(s)$ and $D_C(s)$ with each diagonal entry being a monomial in s. Consequently, VIAP$(A; r)$ reduces to an independent assignment problem [14, Remark 6.2.10], which allows us to state the time complexity of the algorithm for degree of determinant as follows.

LEMMA 5.1. *Let $\tilde{A}(s)$ be an $n \times n$ regular mixed polynomial matrix. If $\tilde{A}(s)$ satisfies* (MP-Q2)*, we obtain an optimal pair for* VIAP$(A; n)$ *in $O(n^4)$ time.*

Proof. Note that the associated LM-polynomial matrix $A(s)$ satisfies (MP-Q2). As an initial B in Step 1 of the algorithm for degree of determinant, we can set $B = \tilde{C}$. In Step 2, E^- can be constructed in $O(n^3)$ time. We can find the shortest path in Step 3 in $O(n^2)$ time. Thus the total complexity of the algorithm for degree of determinant is $O(n^4)$ time. ∎

Lemma 5.1 implies that the time complexity of the algorithm for degrees of all cofactors is $O(n^4)$ as follows.

THEOREM 5.2. *Let $\tilde{A}(s)$ be an $n \times n$ regular mixed polynomial matrix that satisfies* (MP-Q2)*. Then the time complexity of the algorithm for degrees of all cofactors is $O(n^4)$.*

Proof. In Step 3, shortest path distances for all pairs of vertices are computed in $O(n^3)$ time by the Warshall–Floyd method. Hence Lemma 5.1 implies that the total complexity is $O(n^4)$. ∎

Gabow and Xu [6] devised an efficient scaling algorithm for the independent assignment problem. By using this algorithm, the algorithm for degrees of all cofactors can be implemented to run in $O(n^3 \log n \log(nN))$ time, where N denotes the highest degree of all the entries in $\tilde{A}(s)$.

6. Degree matrix. This section presents an algorithm for computing a *degree matrix* defined as follows.

DEFINITION 6.1 (degree matrix). *Let $A(s) = \binom{Q(s)}{T(s)}$ be an $n \times n$ regular LM-polynomial matrix with row set $R = R_Q \cup R_T$ and column set C. Consider another LM-polynomial matrix $A'(s)$ defined by*

$$A'(s) = \begin{array}{c} \\ R_Q \\ R_T \end{array} \begin{array}{c} \begin{array}{cc} C & \hat{C} \end{array} \\ \begin{pmatrix} Q(s) & Q(s) \\ T(s) & O \end{pmatrix} \end{array},$$

where $\hat{C}$ is the copy of C. We denote the copy of $j \in C$ by $\hat{j} \in \hat{C}$. The degree matrix is the matrix $\Theta = (\theta_{kl})$ whose row and column sets are both identical with C such that each entry θ_{kl} is given by $\theta_{kl} = \deg \det A'[R, C \setminus \{l\} \cup \{\hat{k}\}]$.

We now explain the meaning of this degree matrix in the case when $Q(s)$ is a constant matrix Q. For an LM-polynomial matrix $A(s) = \binom{Q}{T(s)}$, consider the following transformation:

$$\begin{pmatrix} S & O \\ O & I_{m_T} \end{pmatrix} \begin{pmatrix} Q \\ T(s) \end{pmatrix}, \tag{6.1}$$

where S is a nonsingular constant matrix and I_{m_T} is the identity matrix of order m_T. The transformation (6.1) does not change the entries in row set R_T and brings an LM-polynomial matrix into another LM-polynomial matrix. By a certain transformation of this type, we obtain an LM-polynomial matrix

$$\check{A}(s) = \begin{matrix} R_Q \\ R_T \end{matrix} \begin{pmatrix} I_{m_Q} & Q' \\ T(s) & \end{pmatrix}.$$

We denote by X the column set of I_{m_Q}. Note that there exists a one-to-one correspondence between $k_Q \in R_Q$ and $l \in X$ with the (k_Q, l) entry of $\check{A}(s)$ being nonzero. The relation between the degree of a cofactor in $\check{A}(s)$ and an entry of the degree matrix Θ is as follows.

LEMMA 6.2. *For any $k_Q \in R_Q$ and $l \in C$, we have $\theta_{kl} = \deg\det \check{A}[R \setminus \{k_Q\}, C \setminus \{l\}]$, where $k \in X$ is the column corresponding to row k_Q.*

Proof. Since we can transform $A(s)$ into $\check{A}(s)$ by row operations, we may assume that Θ is defined in terms of $\check{A}(s)$. Hence we have

$$\theta_{kl} = \deg\det \begin{pmatrix} \check{A}[R_Q, C \setminus \{l\}] & \check{A}[R_Q, \{k\}] \\ \check{A}[R_T, C \setminus \{l\}] & \mathbf{0} \end{pmatrix} = \deg\det \check{A}[R \setminus \{k_Q\}, C \setminus \{l\}],$$

because $\check{A}[R_Q, \{k\}]$ has only one nonzero entry in row k_Q. □

By Lemma 6.2, the entries in row k of Θ coincide with the degrees of cofactors obtained by deleting row k_Q from $\check{A}(s)$.

We now define the following problem associated with an $n \times n$ regular LM-polynomial matrix $A(s) = \binom{Q(s)}{T(s)}$.

[DM$(A; k, l)$] Find a pair (M, B) of a matching $M \subseteq E$ and a base $B \in \mathcal{B}^-$ maximizing $w(M) + \omega^-(B)$ subject to

$$\partial^+ M = R_T, \quad \partial^- M = B \setminus \{l\} \cup \{k\}, \quad l \in B, \quad k \notin B.$$

This problem is similar to VIAP$(A; m_T)$, which can be reformulated as follows.

[VIAP$(A; m_T)$] Find a pair (M, B) of a matching $M \subseteq E$ and a base $B \in \mathcal{B}^-$ maximizing $w(M) + \omega^-(B)$ subject to

$$\partial^+ M = R_T \quad \text{and} \quad \partial^- M = B.$$

See Figure 6.1 for the comparison among feasible solutions (M, B) for VIAP$(A; m_T)$, DOC$(A; k_T, l)$, and DM$(A; k, l)$.

The value of a feasible pair (M, B) for DM$(A; k, l)$ is given by

$$\begin{aligned} w(M) + \omega^-(B) &= w(M) + \omega_Q(C \setminus B) \\ &= \deg\det Q[R_Q, C \setminus B] + \sum_{(i,j) \in M} \deg T_{ij}(s) \\ &= \deg\det A'[R_Q, (C \setminus \{l\} \cup \{\hat{k}\}) \setminus \partial^- M] + \sum_{(i,j) \in M} \deg T_{ij}(s). \end{aligned}$$

Then it follows from Theorem 3.5 that the value of θ_{kl} coincides with the optimal value of DM$(A; k, l)$.

We can find an optimal pair (M, B) for VIAP$(A; m_T)$ by using the algorithm for degree of determinant. We then construct the auxiliary graph G^* with respect

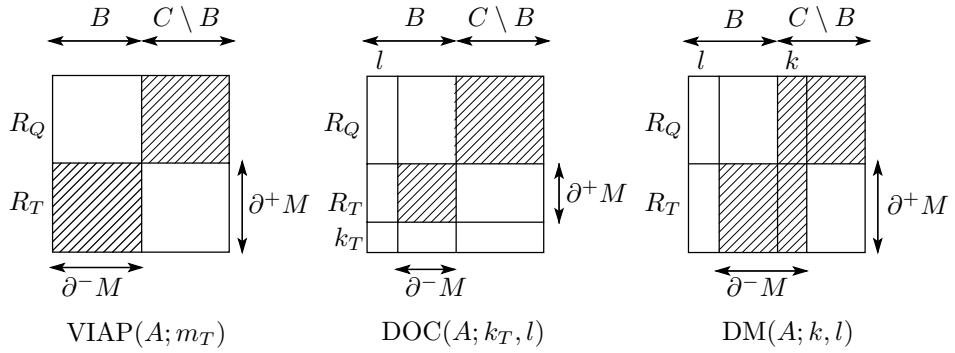


FIG. 6.1. *Comparison among feasible solutions* (M, B) *for* VIAP$(A; m_T)$, DOC$(A; k_T, l)$, *and* DM$(A; k, l)$.

to (M, B). The following theorem leads to an algorithm for computing the degree matrix. The proof is omitted, as it is quite similar to that of Theorem 4.3.

THEOREM 6.3. *Let* (M, B) *be an optimal pair for* VIAP$(A; m_T)$. *For any* $k \in C$ *and* $l \in C$, *we have*

$$\theta_{kl} = \Omega_{m_T}(M, B) - d(l, k),$$

where $d(l, k)$ *denotes the shortest path distance from* l *to* k *with respect to the arc length* γ *in* G^*. □

The algorithm for computing a degree matrix is summarized as follows. The output of this algorithm is a degree matrix $\Theta = (\theta_{kl})$.

ALGORITHM FOR DEGREE MATRIX.

Step 1: Find an optimal pair (M, B) for VIAP$(A; m_T)$ by the algorithm for degree of determinant.

Step 2: Construct an auxiliary graph G^* with respect to (M, B).

Step 3: Compute the shortest path distances for all pairs of $k \in C$ and $l \in C$. For each k and l, set $\theta_{kl} := \Omega_{m_T}(M, B) - d(l, k)$.

Step 4: Return Θ.

The time complexity of the algorithm for degree matrix is the same as that of the algorithm for degree of determinant, because the shortest path distances in Step 3 can be computed in $O(n^3)$ time by the Warshall–Floyd method [5, 16]. For example, if an LM-polynomial matrix $A(s)$ satisfies (MP-Q2), the total running time is $O(n^4)$. If $A(s)$ is a coefficient matrix of circuit equations, the complexity is improved under the genericity assumption that the physical parameters in the constitutive equations are algebraically independent.

THEOREM 6.4. *For a linear time-invariant electric circuit with* n *elements, we denote by* $A(s)$ *a* $2n \times 2n$ *coefficient matrix of circuit equations. Then the algorithm for degree matrix can be implemented to run in* $O(n^3)$ *time if the set of nonzero entries coming from the physical parameters are algebraically independent.*

Proof. Let us denote the row sets of $A(s)$ corresponding to KCL and KVL by R_I and R_V, respectively. We show that the time complexity of the algorithm for degree of determinant is $O(n^3)$. An initial B in Step 1 can be found in $O(n^3)$ time, because $A[R_I \cup R_V, C]$ is a constant matrix. In Step 2, the construction of E^- is as follows. Let B be a base, and Γ be a network graph of the circuit with vertex set W and edge set F. We split $C \setminus B$ into B_I and B_V such that $A[R_I, B_I]$ and $A[R_V, B_V]$ are nonsingular. Let us denote a spanning tree corresponding to B_I in Γ

by T_I, and a cotree corresponding to B_V by $\overline{T}_V$. Consider subgraphs $\Gamma_I = (W, T_I)$ and $\Gamma_V = (W, F \setminus \overline{T}_V)$ of Γ. For each $e = (u, v) \in F \setminus T_I$, we find a path $P_I(e)$ from u to v in Γ_I in $O(n)$ time, because the number of edges is $O(n)$. Similarly, for each $e = (u, v) \in \overline{T}_V$, we find a path $P_V(e)$ from u to v in Γ_V in $O(n)$ time. Then we obtain $E^- = \{(\bar{e}, e) \mid e \in F \setminus T_I, \bar{e} \in P_I(e)\} \cup \{(e, \bar{e}) \mid e \in \overline{T}_V, \bar{e} \in P_V(e)\}$. Thus E^- can be constructed in $O(n^2)$ time. A shortest path in Step 3 can be found in $O(n^2)$ time. Therefore, the time complexity of the algorithm for degree of determinant is $O(n^3)$, which implies that Step 1 of the algorithm for degree matrix requires $O(n^3)$ time.

In Step 3, the Warshall–Floyd method finds the shortest path distances in $O(n^3)$ time. Thus, the total time complexity of the algorithm for degree matrix is $O(n^3)$. □

The notion of the degree matrix plays a key role in the index reduction method for the differential-algebraic equation arising from the hybrid analysis in circuit simulation. Since the LM-polynomial matrix considered there is a coefficient matrix of the circuit equations, the degree matrix can be obtained in $O(n^3)$ time by Theorem 6.4. This improves the time complexity of finding the minimum index hybrid analysis in [9] by a factor of n^3.

REFERENCES

[1] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, SIAM, Philadelphia, 1996.

[2] P. Bujakiewicz, *Maximum Weighted Matching for High Index Differential Algebraic Equations*, Doctor's dissertation, Delft University of Technology, Delft, The Netherlands, 1994.

[3] P. Bujakiewicz and P. P. J. van den Bosch, *Determination of perturbation index of a DAE with maximum weighted matching algorithm*, in Proceedings of the IEEE/IFAC Joint Symposium on Computer-Aided Control System Design, 1994, pp. 129–136.

[4] A. W. M. Dress and W. Wenzel, *Valuated matroids*, Adv. Math., 93 (1992), pp. 214–250.

[5] R. W. Floyd, *Algorithm 97—shortest path*, Communications of the ACM, 5 (1962), p. 345.

[6] H. N. Gabow and Y. Xu, *Efficient theoretic and practical algorithms for linear matroid intersection problems*, J. Comput. System Sci., 53 (1996), pp. 129–147.

[7] M. Iri and N. Tomizawa, *An algorithm for finding an optimal "independent assignment,"* J. Oper. Res. Soc. Japan, 19 (1976), pp. 32–57.

[8] S. Iwata and K. Murota, *Combinatorial relaxation algorithm for mixed polynomial matrices*, Math. Program., 90 (2001), pp. 353–371.

[9] S. Iwata and M. Takamatsu, *Index minimization of differential-algebraic equations in hybrid analysis for circuit simulation*, Math. Program., to appear.

[10] K. Murota, *Use of the concept of physical dimensions in the structural approach to systems analysis*, Japan J. Appl. Math., 2 (1985), pp. 471–494.

[11] K. Murota, *Valuated matroid intersection* I: *Optimality criteria*, SIAM J. Discrete Math., 9 (1996), pp. 545–561.

[12] K. Murota, *Valuated matroid intersection* II: *Algorithms*, SIAM J. Discrete Math., 9 (1996), pp. 562–576.

[13] K. Murota, *On the degree of mixed polynomial matrices*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 196–227.

[14] K. Murota, *Matrices and Matroids for Systems Analysis*, Springer-Verlag, Berlin, 2000.

[15] K. Murota and M. Iri, *Structural solvability of systems of equations—A mathematical formulation for distinguishing accurate and inaccurate numbers in structural analysis of systems*, Japan J. Appl. Math., 2 (1985), pp. 247–271.

[16] S. Warshall, *A theorem on Boolean matrices*, J. ACM, 9 (1962), pp. 11–12.

SIAM J. DISCRETE MATH.
Vol. 23, No. 2, pp. 661–664

YET ANOTHER GENERALIZATION OF POSTNIKOV'S HOOK LENGTH FORMULA FOR BINARY TREES*

GUO-NIU HAN†

Abstract. We discover another one-parameter generalization of Postnikov's hook length formula for binary trees. The particularity of our formula is that the hook length h_v appears as an exponent. As an application, another simple hook length formula for binary trees is derived when the underlying parameter takes the value 1/2.

Key words. hook length, Postnikov's formula, binary tree

AMS subject classifications. 05A15, 05A19, 05C05

DOI. 10.1137/080720498

1. Introduction. Consider the set $\mathcal{B}(n)$ of all binary trees with n vertices. For each vertex v of $T \in \mathcal{B}(n)$, the *hook length* of v, denoted by h_v or just h for short, is the number of descendants of v (including v). The *hook length multiset* of T, denoted by $\mathcal{H}(T)$, is the multiset of all hook lengths of T. The following hook length formula for binary trees

$$\sum_{T\in\mathcal{B}(n)} \prod_{h\in\mathcal{H}(T)} \left(1+\frac{1}{h}\right) = \frac{2^n}{n!}(n+1)^{n-1} \tag{1}$$

was discovered by Postnikov [Po]. Further combinatorial proofs and extensions have been proposed by several authors [CY, GS, MY, Se]. In particular, Lascoux conjectured the following one-parameter generalization:

$$\sum_{T\in\mathcal{B}(n)} \prod_{h\in\mathcal{H}(T)} \left(x+\frac{1}{h}\right) = \frac{1}{(n+1)!}\prod_{k=0}^{n-1}((n+1+k)x+n+1-k), \tag{2}$$

which was, subsequently, proved by Du and Liu [DL]. The latter generalization appears to be very natural, because the *left-hand side* of (2) can be obtained from the left-hand side of (1) by replacing 1 by x.

It is also natural to look for an extension of (1) by introducing a new variable z in the *right-hand side*, namely, by replacing $2^n(n+1)^{n-1}/n!$ by $2^n z(n+z)^{n-1}/n!$. It so happens that the corresponding left-hand side is also a sum on binary trees, but this time the hook length h_v appears as an exponent. The purpose of this note is to prove the following theorem.

THEOREM 1. *For each positive integer n we have*

$$\sum_{T\in\mathcal{B}(n)} \prod_{h\in\mathcal{H}(T)} \frac{(z+h)^{h-1}}{h(2z+h-1)^{h-2}} = \frac{2^n z}{n!}(n+z)^{n-1}. \tag{3}$$

*Received by the editors April 7, 2008; accepted for publication (in revised form) October 13, 2008; published electronically March 4, 2009.

http://www.siam.org/journals/sidma/23-2/72049.html

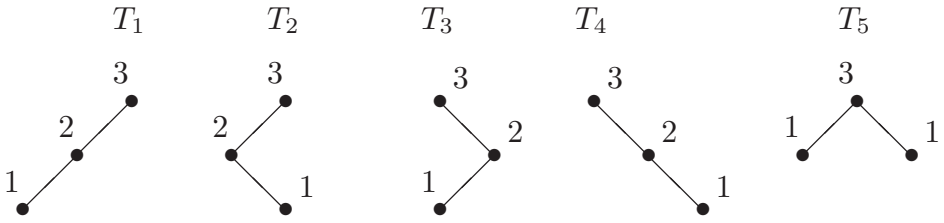
†Center for Combinatorics, LPMC, Nankai University, Tianjin 300071, People's Republic of China, and I.R.M.A. UMR 7501, Université Louis Pasteur et CNRS, 7, rue René-Descartes, F-67084 Strasbourg, France (guoniu@math.u-strasbg.fr).

With $z = 1$ in (3) we recover Postnikov's identity (1). The following corollary is derived from our identity (3) by taking $z = 1/2$.

COROLLARY 2. *For each positive integer n we have*

$$\sum_{T\in\mathcal{B}(n)}\prod_{h\in\mathcal{H}(T)}\left(1+\frac{1}{2h}\right)^{h-1}=\frac{(2n+1)^{n-1}}{n!}. \tag{4}$$

2. Proof of Theorem 1. Let us take an example before proving Theorem 1. There are five binary trees with $n = 3$ vertices, labeled by their hook lengths:



The hook lengths of T_1, T_2, T_3, and T_4 are all the same $1, 2, 3$; but the hook lengths of T_5 are $1, 1, 3$. The left-hand side of (3) is then equal to

$$4\times\frac{1}{(2z)^{-1}}\cdot\frac{(z+2)^1}{2}\cdot\frac{(z+3)^2}{3(2z+1)}+\frac{1}{(2z)^{-1}}\cdot\frac{1}{(2z)^{-1}}\cdot\frac{(z+3)^2}{3(2z+1)}=\frac{2^3z(z+3)^2}{3!}.$$

Let $y(x)$ be a formal power series in x such that

$$y(x)=e^{xy(x)}. \tag{5}$$

By the Lagrange inversion formula $y(x)^z$ has the following explicit expansion:

$$y(x)^z=\sum_{n\geq 0}z(n+z)^{n-1}\frac{x^n}{n!}. \tag{6}$$

Since $y^{2z} = (y^z)^2$, we have

$$\sum_{n\geq 0}2z(n+2z)^{n-1}\frac{x^n}{n!}=\left(\sum_{n\geq 0}z(n+z)^{n-1}\frac{x^n}{n!}\right)^2. \tag{7}$$

Comparing the coefficients of x^n on both sides of (7) yields the following lemma.

LEMMA 3. *We have*

$$\frac{2z(n+2z)^{n-1}}{n!}=\sum_{k=0}^{n}\frac{z(k+z)^{k-1}}{k!}\times\frac{z(n-k+z)^{n-k-1}}{(n-k)!}. \tag{8}$$

In fact, Lemma 3 can be obtained from Abel's celebrated generalization of the binomial formula by a simple change of variables (see [Mo, p. 12] or [Ri, p. 18]).

Proof of Theorem 1. Let

$$P(n)=\sum_{T\in\mathcal{B}(n)}\prod_{h\in\mathcal{H}(T)}\frac{(z+h)^{h-1}}{h(2z+h-1)^{h-2}}.$$

We show that $P(n)$ satisfies a weighted Catalan recurrence (see (9)). In fact, each binary tree T with n vertices is obtained by attaching a left tree and a right tree (with k and $n-k-1$ vertices, respectively) at the root v, which has hook length $h_v = n$. Hence, $P(0) = 1$ and

$$P(n) = \sum_{k=0}^{n-1} P(k)P(n-1-k) \times \frac{(z+n)^{n-1}}{n(2z+n-1)^{n-2}} \quad (n \geq 1). \tag{9}$$

It is routine to verify that $P(n) = 2^n z(z+n)^{n-1}/n!$ for $n = 1, 2, 3$. Suppose that $P(k) = 2^k z(z+k)^{k-1}/k!$ for $k \leq n-1$. From identity (9) and Lemma 3 we have

$$\begin{aligned} P(n) &= \sum_{k=0}^{n-1} \frac{2^k z(z+k)^{k-1}}{k!} \times \frac{2^{n-k-1} z(z+n-k-1)^{n-k-2}}{(n-k-1)!} \times \frac{(z+n)^{n-1}}{n(2z+n-1)^{n-2}} \\ &= \frac{2^n z}{n!}(z+n)^{n-1}. \end{aligned}$$

By induction, (3) is true for any positive integer n. □

3. Conclusion and remarks. The present hook length formula was originally discovered by using the *expansion technique*, developed in [Ha]. A unified formula that includes both the Lascoux–Du–Liu generalization (2) and the present generalization (3) has also been proved in [Ha, Theorem 6.8]. In [Ya] Yang has extended (3) to binomial families of trees.

The right-hand sides of (3) and (4) have been studied by other authors [GS, DL, MY], but our formula has the following two major differences: (i) the hook length h_v appears as an exponent; (ii) the underlying set remains the set of binary trees, whereas in the above-mentioned papers the summation has been changed to the set of m-ary trees or plane forests. It is interesting to compare Corollary 2 with the following results obtained by Du and Liu [DL]. Note that the right-hand sides of (4), (10), and (11) are all identical!

PROPOSITION 4. *For each positive integer n we have*

$$\sum_{T \in \mathcal{T}(n)} \prod_{v \in I(T)} \left(\frac{2}{3} + \frac{1}{3h_v} \right) = \frac{(2n+1)^{n-1}}{n!}, \tag{10}$$

where $\mathcal{T}(n)$ is the set of all 3-ary trees with n internal vertices and $I(T)$ is the set of all internal vertices of T.

PROPOSITION 5. *For each positive integer n we have*

$$\sum_{T \in \mathcal{F}(n)} \prod_{v \in T} \left(2 - \frac{1}{h_v} \right) = \frac{(2n+1)^{n-1}}{n!}, \tag{11}$$

where $\mathcal{F}(n)$ is the set of all plane forests with n vertices.

Acknowledgments. The author thanks the referee and Laura Yang who made knowledgeable remarks that have been taken into account in the final version.

REFERENCES

[CY] W. Chen and L. Yang, *On Postnikov's hook length formula for binary trees,* European J. Combin., 29 (2008), pp. 1563–1565.
[DL] R. Du and F. Liu, (k,m)-*Catalan numbers and hook length polynomials for plane trees,* European J. Combin., 28 (2007), pp. 1312–1321.
[GS] I.M. Gessel and S. Seo, *A refinement of Cayley's formula for trees,* Electron. J. Combin., 11 (2004/06), pp. 27, 23.
[Ha] G.-N. Han, *Discovering hook length formulas by an expansion technique,* Electron., J. Combin., 15 (2008), R133.
[Mo] J.W. Moon, *Counting Labelled Trees,* From lectures delivered to the 12th Biennial Seminar of the Canadian Mathematical Congress (Vancouver, 1969), Canad. Math. Monogr. 1, Canadian Mathematical Congress, Montreal, 1970.
[MY] J.W. Moon and L. Yang, *Postnikov identities and Seo's formulas,* Bull. Inst. Combin. Appl., 49 (2007), pp. 21–31.
[Po] A. Postnikov, *Permutohedra, associahedra, and beyond,* arXiv:math. CO/0507163, 2004.
[Ri] J. Riordan, *Combinatorial Identities,* John Wiley & Sons, New York, 1968.
[Se] S. Seo, *A combinatorial proof of Postnikov's identity and a generalized enumeration of labeled trees,* Electron. J. Combin., 11 (2004/06), p. 9.
[Ya] L. Yang, *Generalizations of Han's hook length identities,* arXiv: 0805. 0109 [math.CO], 2008.

SIAM J. DISCRETE MATH.
Vol. 23, No. 2, pp. 665–679

ON THE INTEGRALITY OF SOME FACILITY LOCATION POLYTOPES*

MOURAD BAÏOU† AND FRANCISCO BARAHONA‡

Abstract. We study a system of linear inequalities associated with some facility location problems. We show that this system defines a polytope with integer extreme points if and only if the graph does not contain a certain type of odd cycles. We also derive odd cycle inequalities and give a separation algorithm.

Key words. facility location, odd cycle inequalities

AMS subject classifications. 05C85, 90C27

DOI. 10.1137/070706070

1. Introduction. Let $G = (V, A)$ be a directed graph, not necessarily connected, where each arc and each node has weight associated with it. We study a "prize collecting" version of a *location problem* (LP) as follows. A set of nodes is selected, usually called *centers*, and then each nonselected node can be assigned to a center. The weight of a node is the revenue obtained by opening a facility at that location, minus the cost of building the facility. The weight of an arc (i, j) is the revenue obtained by assigning the location i to the location j, minus the cost originated by this assignment. The goal is to maximize the sum of the weights of the selected nodes plus the sum of the weights yielded by the assignment. The linear system below defines a linear programming relaxation:

$$\max \sum w(u,v)x(u,v) + \sum w(v)y(v)$$

$$\sum_{(u,v)\in A} x(u,v) + y(u) \le 1 \quad \forall u \in V, \tag{1}$$

$$x(u,v) \le y(v) \quad \forall (u,v) \in A, \tag{2}$$

$$0 \le y(v) \le 1 \quad \forall v \in V, \tag{3}$$

$$x(u,v) \ge 0 \quad \forall (u,v) \in A. \tag{4}$$

For each node u, the variable $y(u)$ takes the value 1 if the node u is selected and 0 otherwise. For each arc (u, v) the variable $x(u, v)$ takes the value 1 if u is assigned to v and 0 otherwise. Inequalities (1) express the fact that either node u can be selected or it can be assigned to another node. Inequalities (2) indicate that if a node u is assigned to a node v, then this last node should be selected. The set of integer vectors that satisfy (1)–(4) corresponds to a *transitive packing* as defined in [15].

Let $P(G)$ be the polytope defined by (1)–(4), and let $LP(G)$ be the convex hull of $P(G) \cap \{0, 1\}^{|V|+|A|}$. Clearly

$$LP(G) \subseteq P(G).$$

*Received by the editors October 23, 2007; accepted for publication (in revised form) November 19, 2008; published electronically March 4, 2009.
http://www.siam.org/journals/sidma/23-2/70607.html
†CNRS, Laboratoire LIMOS, Campus des Cézeaux BP 125, 63173 Aubière cedex, France (baiou@isima.fr).
‡IBM T. J. Watson Research Center, Yorktown Heights, NY 10589 (barahon@us.ibm.com).

In this paper we characterize the graphs G for which $LP(G) = P(G)$. More precisely, we show that $LP(G) = P(G)$ if and only if G does not contain certain types of "odd" cycles. We also give a polynomial algorithm to recognize the graphs in this class.

The *uncapacitated facility location problem* (UFLP) is a variation where V is partitioned into V_1 and V_2. The set V_1 corresponds to the customers, and the set V_2 corresponds to the potential facilities. Each customer in V_1 should be assigned to an opened facility in V_2. This is obtained by considering $A \subseteq V_1 \times V_2$, fixing to zero the variables y for the nodes in V_1, and setting into equations all the inequalities (1) for the nodes in V_1. More precisely, the linear programming relaxation for this case is

$$\min \sum c(u,v)x(u,v) + \sum d(v)y(v)$$

$$\sum_{(u,v)\in A} x(u,v) = 1 \quad \forall u \in V_1, \tag{5}$$

$$x(u,v) \le y(v) \quad \forall (u,v) \in A, \tag{6}$$

$$0 \le y(v) \le 1 \quad \forall v \in V_2, \tag{7}$$

$$x(u,v) \ge 0 \quad \forall (u,v) \in A. \tag{8}$$

Here we also characterize the cases for which (5)–(8) define an integral polytope.

The facets of the uncapacitated facility location polytope have been studied in [13], [11], [5], [6], [3]. In [1] we gave a description of $LP(G)$ for Y-free graphs. The UFLP has also been studied from the point of view of approximation algorithms in [16], [7], [17], [2], [18], and others. Other references on this problem are [10] and [14]. The relationship between location polytopes and the stable set polytope has been studied in [11], [5], [6], [12], and others.

For a directed graph $G = (V, A)$ and a set $W \subset V$, we denote by $\delta^+(W)$ the set of arcs $(u, v) \in A$, with $u \in W$ and $v \in V \setminus W$. Also, we denote by $\delta^-(W)$ the set of arcs (u, v), with $v \in W$ and $u \in V \setminus W$. We write $\delta^+(v)$ and $\delta^-(v)$ instead of $\delta^+(\{v\})$ and $\delta^-(\{v\})$, respectively. If there is a risk of confusion, we use δ^+_G and δ^-_G. A node u with $\delta^+(u) = \emptyset$ is called a *pendent* node.

A simple cycle C is an ordered sequence

$$v_0, a_0, v_1, a_1, \ldots, a_{p-1}, v_p,$$

where

- v_i, $0 \le i \le p-1$, are distinct nodes,
- a_i, $0 \le i \le p-1$, are distinct arcs,
- either v_i is the tail of a_i and v_{i+1} is the head of a_i, or v_i is the head of a_i and v_{i+1} is the tail of a_i for $0 \le i \le p-1$, and
- $v_0 = v_p$.

By setting $a_p = a_0$, we associate with C three more sets as below.

- We denote by $\hat{C}$ the set of nodes v_i, such that v_i is the head of a_{i-1} and also the head of a_i, $1 \le i \le p$.
- We denote by $\dot{C}$ the set of nodes v_i, such that v_i is the tail of a_{i-1} and also the tail of a_i, $1 \le i \le p$.
- We denote by $\tilde{C}$ the set of nodes v_i, such that either v_i is the head of a_{i-1} and also the tail of a_i, or v_i is the tail of a_{i-1} and also the head of a_i, $1 \le i \le p$.

Notice that $|\hat{C}| = |\dot{C}|$. A cycle will be called *odd* if $p + |\dot{C}|$ (or $|\tilde{C}| + |\dot{C}|$) is odd; otherwise it will be called *even*. A cycle C with $\dot{C} = \emptyset$ is a *directed* cycle. The set of arcs in C is denoted by $A(C)$. We plan to prove that $LP(G) = P(G)$ if and only if G has no odd cycle.

If we do not require $v_0 = v_p$, we have a *path* P. In a similar way we define $\dot{P}$, $\hat{P}$, and $\tilde{P}$, excluding v_0 and v_p. We say that P is *odd* if $p + |\dot{P}|$ is odd; otherwise it is *even*. For the path P, the nodes $v_1, \ldots, v_{p-1}$ are called *internal*.

If G is a connected graph and there is a node u such that its removal disconnects G, we say that u is an *articulation point*. A graph is said to be *two-connected* if at least two nodes should be removed to disconnect it. For simplicity, sometimes we use z to denote the vector (x, y), i.e., $z(u) = y(u)$ and $z(u, v) = x(u, v)$. Also for $S \subseteq V \cup A$ we use $z(S)$ to denote $z(S) = \sum_{a \in S} z(a)$.

A *polyhedron* P is defined by a set of linear inequalities, i.e., $P = \{x \mid Ax \leq b\}$. A *face* of P is obtained by setting into equations some of these inequalities. An *extreme point* of P is given by a face that contains a unique element. In other words, some inequalities are set to equations so that this system has a unique solution. A polyhedron whose extreme points are integer is called an *integral polyhedron*.

This paper is organized as follows. In section 2 we give a decomposition theorem that shows that one has to concentrate on two-connected graphs. In section 3 we describe some transformations of the graph that are needed in the following section. Section 4 is devoted to two-connected graphs. In section 5 we study graphs with odd cycles. The separation problem for the so-called odd cycle inequalities is studied in section 6. In section 7 we show how to test the existence of an odd cycle. Section 8 is devoted to the bipartite case.

2. Decomposition. In this section we consider a graph $G = (V, A)$ that decomposes into two graphs $G_1 = (V_1, A_1)$ and $G_2 = (V_2, A_2)$, with $V = V_1 \cup V_2$, $V_1 \cap V_2 = \{u\}$, $A = A_1 \cup A_2$, $A_1 \cap A_2 = \emptyset$. We define G'_1, which is obtained from G_1 after replacing u by u'. We also define G'_2, which is obtained from G_2 after replacing u by u''. See Figure 1. The theorem below shows that we have to concentrate on two-connected graphs.

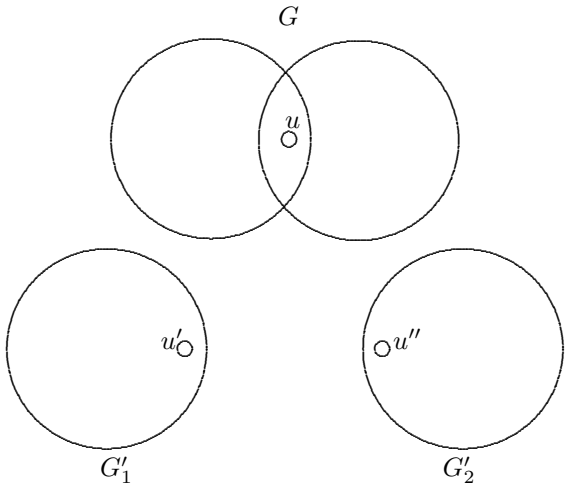


FIG. 1.

THEOREM 1. *Suppose that the system*

$$Az' \leq b, \tag{9}$$

$$z'\left(\delta^+_{G'_1}(u')\right) + z'(u') \leq 1 \tag{10}$$

describes $LP(G'_1)$. *Suppose that* (9) *contains the inequalities* (1)–(4) *except for* (10). *Similarly, suppose that*

$$Cz'' \leq d, \tag{11}$$

$$z''\Big(\delta^+_{G'_2}(u'')\Big) + z''(u'') \leq 1 \tag{12}$$

describes $LP(G'_2)$. *Also* (11) *contains the inequalities* (1)–(4) *except for* (12). *Then the system below describes an integral polyhedron:*

$$Az' \leq b, \tag{13}$$

$$Cz'' \leq d, \tag{14}$$

$$z'\Big(\delta^+_{G'_1}(u')\Big) + z''\Big(\delta^+_{G'_2}(u'')\Big) + z'(u') \leq 1, \tag{15}$$

$$z'(u') = z''(u''). \tag{16}$$

Proof. Let $(\bar{z}', \bar{z}'')$ be an extreme point of the polytope defined by the above system. We study two cases.

Case 1. $\bar{z}'(u') = 0$. We have that $\bar{z}' \in LP(G'_1)$ and $\bar{z}'' \in LP(G'_2)$. If $\bar{z}'$ is an extreme point of $LP(G'_1)$, we have to consider two subcases:

- $\bar{z}'\big(\delta^+_{G'_1}(u')\big) = 0$. If $\bar{z}''$ is not an extreme point of $LP(G'_2)$, $\bar{z}'' = 1/2\lambda_1 + 1/2\lambda_2$, with λ_1, λ_2 in $LP(G'_2)$, $\lambda_1 \neq \lambda_2$. Since $\lambda_1\big(\delta^+_{G'_2}(u'')\big) \leq 1$, $\lambda_2\big(\delta^+_{G'_2}(u'')\big) \leq 1$, we have that $(\bar{z}', \bar{z}'') = 1/2(\bar{z}', \lambda_1) + 1/2(\bar{z}', \lambda_2)$, with $(\bar{z}', \lambda_1)$ and $(\bar{z}', \lambda_2)$ satisfying (13)–(16), which is a contradiction. Thus $\bar{z}''$ is an extreme point and $(\bar{z}', \bar{z}'')$ is an integral vector.
- $\bar{z}'\big(\delta^+_{G'_1}(u')\big) = 1$. This implies that $\bar{z}''\big(\delta^+_{G'_2}(u'')\big) = 0$. If $\bar{z}''$ is not an extreme point, then $\bar{z}'' = 1/2\lambda_1 + 1/2\lambda_2$, with λ_1, λ_2 in $LP(G'_2)$, $\lambda_1 \neq \lambda_2$. Since $\lambda_1\big(\delta^+_{G'_2}(u'')\big) = 0 = \lambda_2\big(\delta^+_{G'_2}(u'')\big)$, we have that $(\bar{z}', \bar{z}'') = 1/2(\bar{z}', \lambda_1) + 1/2(\bar{z}', \lambda_2)$, with $(\bar{z}', \lambda_1)$ and $(\bar{z}', \lambda_2)$ satisfying (13)–(16), which is a contradiction. Thus $\bar{z}''$ is an extreme point and $(\bar{z}', \bar{z}'')$ is an integral vector.

Now we should study the situation in which $\bar{z}'$ and $\bar{z}''$ are not extreme points.

We should have $\bar{z}' = 1/2\omega_1 + 1/2\omega_2$, with ω_1, ω_2 in $LP(G'_1)$, $\omega_1 \neq \omega_2$. If $\omega_1\big(\delta^+_{G'_1}(u')\big) = \omega_2\big(\delta^+_{G'_1}(u')\big) = \bar{z}'\big(\delta^+_{G'_1}(u')\big)$, we have $(\bar{z}', \bar{z}'') = 1/2(\omega_1, \bar{z}'') + 1/2(\omega_2, \bar{z}'')$, with $(\omega_1, \bar{z}'')$ and $(\omega_2, \bar{z}'')$ satisfying (13)–(16), which is a contradiction.

Now we assume that

$$\omega_1\Big(\delta^+_{G'_1}(u')\Big) = \bar{z}'\Big(\delta^+_{G'_1}(u')\Big) - \epsilon,$$
$$\omega_2\Big(\delta^+_{G'_1}(u')\Big) = \bar{z}'\Big(\delta^+_{G'_1}(u')\Big) + \epsilon,$$

with $\epsilon > 0$.

We also have $\bar{z}'' = 1/2\lambda_1 + 1/2\lambda_2$, with λ_1, λ_2 in $LP(G'_2)$, $\lambda_1 \neq \lambda_2$. If $\lambda_1\big(\delta^+_{G'_2}(u'')\big) = \lambda_2\big(\delta^+_{G'_2}(u'')\big) = \bar{z}''\big(\delta^+_{G'_2}(u'')\big)$, we obtain a contradiction as above. Thus we suppose that

$$\lambda_1\Big(\delta^+_{G'_2}(u'')\Big) = \bar{z}''\Big(\delta^+_{G'_2}(u'')\Big) + \rho,$$
$$\lambda_2\Big(\delta^+_{G'_2}(u'')\Big) = \bar{z}''\Big(\delta^+_{G'_2}(u'')\Big) - \rho,$$

with $\rho > 0$.

We can assume that $\epsilon = \rho$; otherwise we can change λ_1 and λ_2. Thus we have $(\bar{z}', \bar{z}'') = 1/2(\omega_1, \lambda_1) + 1/2(\omega_2, \lambda_2)$, with (ω_1, λ_1) and (ω_2, λ_2) satisfying (13)–(16), which is a contradiction.

Case 2. $0 < \bar{z}'(u')$. We have that $\bar{z}' \in LP(G_1')$ and $\bar{z}'' \in LP(G_2')$. Thus $\bar{z}'$ is a convex combination of extreme points μ_i of $LP(G_1')$ that satisfy with equality every constraint that is satisfied with equality by $\bar{z}'$. Also $\bar{z}''$ is a convex combination of extreme points ϕ_j of $LP(G_2')$ that satisfy with equality every constraint satisfied with equality by $\bar{z}''$.

We can assume that $\mu_1(u') = 1 = \phi_1(u'')$. After putting together these two vectors we obtain a 0-1 vector that satisfies with equality every constraint that is satisfied with equality by the original vector $(\bar{z}', \bar{z}'')$, which is a contradiction. □

We have the following corollary.

COROLLARY 2. *The polytope $LP(G)$ is defined by the system* (13)–(16) *after identifying the variables $z'(u')$ and $z''(u'')$.*

This last corollary shows that if $LP(G_1')$ and $LP(G_2')$ are defined by (1)–(4), then $LP(G)$ is also defined by (1)–(4). Thus we have to concentrate on graphs that are two-connected. A result analogous to Theorem 1, for the stable set polytope, has been given in [8].

3. Graph transformations. First we plan to prove that if G has no odd cycle, then $LP(G) = P(G)$. The proof consists of assuming that $\bar{z}$ is a fractional extreme point of $P(G)$ and arriving at a contradiction. Below we give several assumptions that can be made about $\bar{z}$ and G; they will be used in the next section.

LEMMA 1. *We can assume that $\bar{z}(u, v) > 0$ for all $(u, v) \in A$.*

Proof. Let G' be the graph obtained after removing all arcs (u, v) with $\bar{z}(u, v) = 0$, and let z' be the vector obtained after removing all components $\bar{z}(u, v) = 0$. Then z' is a fractional extreme point of $P(G')$. □

LEMMA 2. *If $0 < \bar{z}(u, v) < \bar{z}(v)$, we can assume that v is a pendent node with $|\delta^-(v)| = 1$ and $\bar{z}(v) = 1$.*

Proof. If v is not pendent or $|\delta^-(v)| > 1$, we can remove (u, v) and add a new node v' and the arc (u, v'). Then we can define $z'(u, v') = \bar{z}(u, v)$, $z'(v') = 1$, and $z'(s, t) = \bar{z}(s, t)$, $z'(r) = z(r)$ for all other nodes and arcs. Let G' be this new graph. We have that the constraints that are tight for $\bar{z}$ are also tight for z', so z' is an extreme point of $P(G')$. □

LEMMA 3. *We can assume that G consists of only one connected component.*

Proof. Let G_1 be a connected component of G. Let z_1 be the projection of $\bar{z}$ onto the space associated with G_1. Then z_1 is an extreme point of $P(G_1)$. □

LEMMA 4. *We can assume that $0 < \bar{z}(u, v) < 1$ for all $(u, v) \in A$.*

Proof. If $\bar{z}(u, v) = 1$, it follows from Lemma 1 that $\delta^-(u) = \emptyset$ and $\delta^+(u) = \{(u, v)\}$. Since $\bar{z}(v) = 1$, Lemma 1 implies that v is pendent. It follows from Lemma 2 that $\bar{z}(r, v) = 1$ for all $(r, v) \in \delta^-(v)$. Therefore, the graph induced by $\delta^-(v)$ is a connected component of G. All variables associated with this connected component take integer values. □

LEMMA 5. *We can assume that either G is two-connected or it consists of a single arc.*

Proof. If G has an articulation point, we can apply Theorem 1 to decompose G into G_1 and G_2. If inequalities (1)–(4) define $LP(G_1)$ and $LP(G_2)$, then a similar system should define $LP(G)$. One can keep decomposing as long as the graph has an articulation point. □

If the graph G consists of a single arc, it is fairly easy to see that $LP(G) = P(G)$, so now we have to deal with the two-connected components. This is treated in the next section.

4. Treating two-connected graphs. In this section we assume that the graph G is two-connected and it has no odd cycle. Let $\bar{z}$ be a fractional extreme point of $P(G)$; we are going to assign labels l to the nodes and arcs and define $z'(u,v) = \bar{z}(u,v) + l(u,v)\epsilon$, $z'(u) = \bar{z}(u) + l(u)\epsilon$, $\epsilon > 0$, for each arc (u,v) and each node u. We shall see that every constraint that is satisfied with equality by $\bar{z}$ is also satisfied with equality by z'. This is the required contradiction.

Given a path $P = v_0, a_0, \ldots, a_{p-1}, v_p$, assume that the label of a_0, $l(a_0)$, has the value 1 or -1. We define the *labeling procedure* as follows.

For $i = 1$ to $p-1$ do the following:

- If v_i is the head of a_{i-1} and it is the tail of a_i, then $l(v_i) = l(a_{i-1})$, $l(a_i) = -l(v_i)$.
- If v_i is the head of a_{i-1} and it is the head of a_i, then $l(v_i) = l(a_{i-1})$, $l(a_i) = l(v_i)$.
- If v_i is the tail of a_{i-1} and it is the head of a_i, then $l(v_i) = -l(a_{i-1})$, $l(a_i) = l(v_i)$.
- If v_i is the tail of a_{i-1} and it is the tail of a_i, then $l(v_i) = 0$, $l(a_i) = -l(a_{i-1})$.

Notice that the labels of v_0 and v_p were not defined.

This procedure will be used in four different cases as below.

Case 1. G contains a directed cycle $C = v_0, a_0, \ldots, a_{p-1}, v_p$. Assume that the head of a_0 is v_1, set $l(v_0) = -1$ and $l(a_0) = 1$, and extend the labels as above.

Case 2. G contains a cycle $C = v_0, a_0, \ldots, a_{p-1}, v_p$ and $\dot{C} \neq \emptyset$. Assume $v_0 \in \dot{C}$. Set $l(v_0) = 0$ and $l(a_0) = 1$, and extend the labels.

The lemma below is needed to show that for v_0, the constraints that were satisfied with equality by $\bar{z}$ remain satisfied with equality.

LEMMA 6. *After labeling as in Cases* 1 *and* 2*, we have* $l(a_{p-1}) = -l(a_0)$.

Proof. Case 1 should be clear, so we have to study Case 2. Let $v_{j(0)}, v_{j(1)}, \ldots, v_{j(k)}$ be the ordered sequence of nodes in $\dot{C}$, with $v_{j(0)} = v_{j(k)}$. A path in C

$$v_{j(i)}, a_{j(i)}, \ldots, a_{j(i+1)-1}, v_{j(i+1)}$$

from $v_{j(i)}$ to $v_{j(i+1)}$ will be called a *segment* and denoted by S_i. A segment is *odd* (resp., *even*) if it contains an *odd* (resp., *even*) number of arcs. Let n_e be the number of even segments and n_o the number of odd segments. We have that $n_e + n_o = |\dot{C}|$. We also have that the parity of p is equal to the parity of n_o. Therefore, $n_o + |\dot{C}|$ should be even.

The labeling has the following properties:

(a) If the segment is odd, then $l(a_{j(i)}) = -l(a_{j(i+1)-1})$.

(b) If the segment is even, then $l(a_{j(i)}) = l(a_{j(i+1)-1})$.

Now we build an undirected cycle as follows. For every node $v_{j(i)}$ we have two nodes u_i^1 and u_i^2; we add an edge between them marked "blue." For every segment from $v_{j(i)}$ to $v_{j(i+1)}$ we have an edge from u_i^2 to u_{i+1}^1. If the segment is odd, we mark the edge "blue"; otherwise we mark it "green." Start by giving the label $l(u_0^2) = 1$ to u_0^2. Continue labeling so that if st is a blue edge, then $l(t) = -l(s)$, and if the edge is green, then $l(t) = l(s)$. The label of u_i^2 corresponds to the label of $a_{j(i)}$, and the label of u_{i+1}^1 corresponds to the label of $a_{j(i+1)-1}$. There is an even number of blue edges in the cycle; therefore, $l(u_0^1) = -l(u_0^2)$. Thus

$$l(a_{p-1}) = -l(a_0). \qquad \square$$

Notice that after the first cycle has been labeled as in Cases 1 or 2, the properties below hold. We shall see that these properties hold throughout the entire labeling procedure.

Property 1. If a node has a nonzero label, then it is the tail of at most one labeled arc.

Property 2. If a node has a zero label, then it is the tail of exactly two arcs with opposite labels, and it is not the head of any labeled arc.

The lemma below shows a property of the labeling procedure that will be used in the analysis of the next case.

LEMMA 7. *Let $P = v_0, a_0, v_1, a_1, \ldots, a_{p-1}, v_p$ be a path. Suppose that we set $l(a_0)$ and we extend the labels; then the label of a_{p-1} is determined by*

- *the orientation of a_0,*
- *the orientation of a_{p-1}, and*
- *the parity of P.*

Proof. Add a node t and the arcs $\bar{a} = (t, v_0)$ and $\tilde{a} = (t, v_p)$ to create a cycle. If the cycle is odd, subdivide $\tilde{a}$ to make it even. Set $l(t) = 0$ and $l(\bar{a}) = 1$, and extend the labels as in Case 2. It follows from Lemma 6 that the label of the arc before $\bar{a}$ is $-l(\bar{a})$; this determines the label of the previous arc, and so on. □

Once a cycle C has been labeled as in Cases 1 or 2, we have to extend the labeling as follows.

Case 3. Suppose that $l(v_0) \neq 0$ for $v_0 \in C$ (v_0 is the head of a labeled arc) and there is a path $P = v_0, a_0, v_1, a_1, \ldots, a_{p-1}, v_p$ in G such that

- v_0 is the head of a_0,
- $v_p \in C$, and
- $\{v_1, \ldots, v_{p-1}\}$ is disjoint from C.

We set $l(a_0) = l(v_0)$ and extend the labels. Case 3 is needed so that any inequality (2) associated with v_0 that is satisfied with equality remains satisfied with equality.

We have to see that the label $l(a_{p-1})$ is such that constraints associated with v_p, that were satisfied with equality, remain satisfied with equality. This is discussed in the next lemma.

LEMMA 8. *If v_p is the head of a_{p-1}, then $l(a_{p-1}) = l(v_p)$. If v_p is the tail of a_{p-1}, then $l(a_{p-1}) = -l(v_p)$.*

Proof. Notice that $v_0 \notin \dot{C}$. In Figure 2 we represent the possible configurations for the paths in C between v_0 and v_p. In this figure we show whether v_0 and v_p are the head or the tail of the arcs in C incident to them. These two paths are denoted by P_1 and P_2. Lemma 7 shows that we have to pay attention to their parity and to the orientation of the first and last arcs.

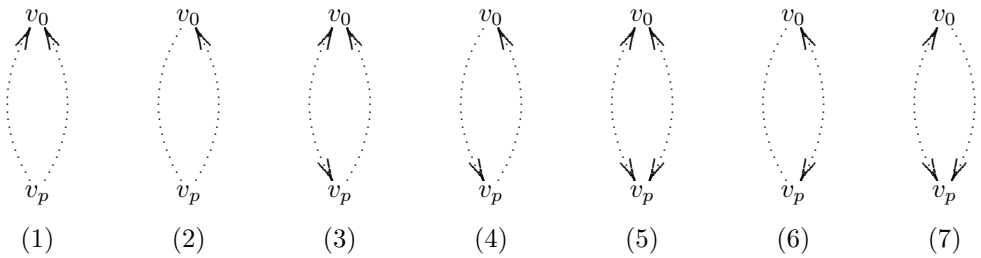


FIG. 2. *Possible paths in C between v_0 and v_p. It is shown whether v_0 and v_p are the head or the tail of the arcs in C incident to them.*

Consider configuration (1); these two paths should have different parity. When adding the path P, an odd cycle is created with either P_1 or P_2. So configuration (1) will not occur. The same happens with configuration (2).

Now we discuss configuration (3). These two paths should have the same parity. If v_p is the tail of a_{p-1}, then P creates an odd cycle with either P_1 or P_2. If v_p is the head of a_{p-1}, then P should have the same parity as P_1 and P_2. Then $l(a_{p-1}) = l(v_p)$.

The study of configuration (4) is similar. The two paths should have the same parity. If v_p is the tail of a_{p-1}, then P creates an odd cycle with either P_1 or P_2. If v_p is the head of a_{p-1}, then P should have the same parity as P_1 and P_2, and $l(a_{p-1}) = l(v_p)$.

For configuration (5), again the two paths should have the same parity. If v_p is the head of a_{p-1}, then P should have the same parity as P_1 and P_2, and $l(a_{p-1}) = l(v_p)$. If v_p is the tail of a_{p-1}, then P should have the same parity as P_1 and P_2, and $l(a_{p-1}) = -l(v_p)$.

Also, in configuration (6) the paths P_1 and P_2 should have the same parity. If v_p is the tail of a_{p-1}, then P forms an odd cycle with either P_1 or P_2. If v_p is the head of a_{p-1}, then P should have the same parity as P_1 and P_2, and $l(a_{p-1}) = l(v_p)$.

In configuration (7), the two paths should also have the same parity. If v_p is the head of a_{p-1}, then P should have the same parity as P_1 and P_2, and $l(a_{p-1}) = l(v_p)$. If v_p is the tail of a_{p-1}, then P should have the same parity as P_1 and P_2, and $l(a_{p-1}) = -l(v_p)$. □

Based on this the labels are extended recursively. Denote by G_l the subgraph defined by the labeled arcs. This is a two-connected graph, so for any two nodes v_0 and v_p it contains a cycle going through these two nodes. Thus we can check if Case 3 applies and extend the labels adding a path to the graph G_l each time. The two lemmas below show that Properties 1 and 2 remain satisfied.

LEMMA 9. *Suppose that v_p has a label different from 0. If v_p is the tail of an arc in G_l, then in Case 3 it cannot be the tail of a_{p-1}. Thus Property 1 remains satisfied.*

Proof. There is a cycle C in G_l containing v_0 and v_p. Property 1 implies that v_0 is the head of at least one arc in C. We can assume that v_p is the tail of an arc in C. Suppose not; let a be an arc in G_l whose tail is v_p. Let u be the head of a. Since G_l is two-connected, there is a path Q from u to a node v in C with $v \neq v_p$. The path Q intersects C only at the node v. We can add a and Q to C and remove the path in C from v_p to v that does not contain v_0 as an internal node.

The cycle C can contain configurations (3), (4), and (6) of Figure 2. In these three cases, the head of a_{p-1} is v_p. □

LEMMA 10. *Let w be a node in G_l with $l(w) = 0$; then in Case 3 we have that $v_p \neq w$. Therefore, Property 2 remains satisfied.*

Proof. Let a_1, a_2 be the two arcs in G_l having w as their tail. If $v_p = w$, the cycle C in Case 3 must contain both arcs a_1 and a_2. But configurations (1) and (2) cannot occur. □

Once Case 3 has been exhausted we might have some nodes in G_l that are only the heads of labeled arcs. For such nodes we have to ensure that inequalities (1) that were satisfied as equalities remain satisfied as equalities. This is treated as follows.

Case 4. Suppose that v_0 is only the head of labeled arcs, and v_0 is not pendent. Then there is a cycle C in G_l and there is a path $P = v_0, a_0, v_1, a_1, \ldots, a_{p-1}, v_p$ in G such that

- $v_0 \in C$ is the tail of a_0,
- $v_p \in C$, and
- $\{v_1, \ldots, v_{p-1}\}$ is disjoint from G_l.

We set $l(a_0) = -l(v_0)$ and extend the labels. We have to see that the label $l(a_{p-1})$ is such that constraints associated with v_p, that were satisfied with equality, remain satisfied with equality. This is discussed below.

LEMMA 11. *In Case* 4 *we have that* v_p *is the tail of* a_{p-1} *and* $l(a_{p-1}) = -l(v_p)$. *Also Properties* 1 *and* 2 *continue to hold.*

Proof. The cycle C can correspond to configurations (1), (3), or (5) of Figure 2.

For configuration (1), the paths P_1 and P_2 have different parities, therefore adding the path P would create an odd cycle.

Consider now configuration (3). The paths P_1 and P_2 have the same parity. If v_p is the tail of a_{p-1}, then adding P to C would create an odd cycle. If v_p is the head of a_{p-1}, we would have a situation treated in Case 3 and configuration (7).

Finally consider configuration (5). If v_p is the head of a_{p-1}, we have a situation treated in Case 3 and configuration (5). If v_p is the tail of a_{p-1}, then P should have the same parity as P_1 and P_2; thus $l(a_{p-1}) = -l(v_p)$. If v_p were the tail of an arc in G_l, we would have a cycle like in configuration (3). Adding P to this cycle would create an odd cycle. Therefore, v_p was not the tail of an arc in G_l and Properties 1 and 2 continue to hold. □

To summarize, the labeling algorithm consists of the following steps.

- Step 1. Identify a cycle C in G and treat it as in Cases 1 or 2. Set $G_l = C$.
- Step 2. For as long as needed, label as in Case 3. Each time add to G_l the new set of labeled nodes and arcs.
- Step 3. If needed, label as in Case 4. Each time add to G_l the new set of labeled nodes and arcs. If some new labels have been assigned in this step, go to Step 2; otherwise stop.

At this point we can discuss the properties of the labeling procedure. The labels are such that any inequality (2) that was satisfied with equality by $\bar{z}$ is also satisfied with equality by z'. To see that inequalities (1) that were tight remain tight, we use Properties 1 and 2:

- Any node that has a nonzero label is the tail of exactly one labeled arc having the opposite label.
- If u is a node with $l(u) = 0$, then there are exactly two labeled arcs having opposite labels and whose tails are u.

Finally, we give the label "0" to all nodes and arcs that are unlabeled; this completes the definition of z'. Lemma 4 shows that inequalities (4) will not be violated. Since nodes v with $\bar{z}(v) = 0$ receive a zero label, and there are no nodes v with $\bar{z}(v) = 1$, we have that inequalities (3) cannot not be violated. Any constraint that is satisfied with equality by $\bar{z}$ is also satisfied with equality by z'. This contradicts the assumption that $\bar{z}$ is an extreme point. We can now state the main result of this section.

THEOREM 3. *If the graph* G *is two-connected and has no odd cycle, then* $LP(G) = P(G)$.

This implies the following.

THEOREM 4. *If* G *is a graph with no odd cycle, then* $LP(G) = P(G)$.

THEOREM 5. *For graphs with no odd cycle, the UFLP is polynomially solvable.*

In some cases one might want to fix to zero the variables y for some set of nodes and also set to equations some of the inequalities (1). This defines a face $Q(G)$ of $P(G)$. We have the following corollary that will be used in section 8.

COROLLARY 6. *If* G *is a graph with no odd cycle, then* $Q(G)$ *is an integral polytope.*

5. Odd cycles. In this section we study the effect of odd cycles in $P(G)$. Let C be an odd cycle. We can define a fractional vector $(\bar{x}, \bar{y}) \in P(G)$ as follows:

$$\bar{y}(u) = 0 \quad \forall \text{ nodes } u \in \dot{C}, \tag{17}$$

$$\bar{y}(u) = 1/2 \quad \forall \text{ nodes } u \in C \setminus \dot{C}, \tag{18}$$

$$\bar{x}(a) = 1/2 \quad \text{for } a \in A(C), \tag{19}$$

$$\bar{y}(v) = 0 \quad \forall \text{ other nodes } v \notin C, \tag{20}$$

$$\bar{x}(a) = 0 \quad \forall \text{ other arcs.} \tag{21}$$

In Figure 3 we show two examples. The numbers close to the nodes correspond to the y variables, and the numbers close to the arcs correspond to the x variables.

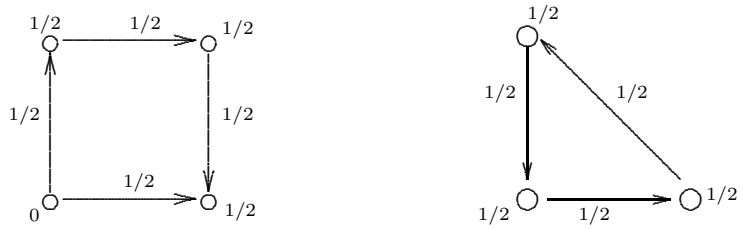


FIG. 3. *Fractional vectors associated with odd cycles.*

Below we show a family of inequalities that separate the vectors defined above from $LP(G)$. We call them *odd cycle* inequalities.

LEMMA 12. *The following inequalities are valid for $LP(G)$:*

$$\sum_{a \in A(C)} x(a) - \sum_{v \in \hat{C}} y(v) \leq \frac{|\tilde{C}| + |\hat{C}| - 1}{2} \tag{22}$$

for every odd cycle C.

Proof. From inequalities (1)–(4) we obtain

$$x(u,v) + x(\delta^+(v)) \leq 1 \quad \text{for every arc } (u,v) \in C,\ v \notin \hat{C},$$
$$x(u,v) - y(v) \leq 0 \quad \text{for every arc } (u,v) \in C,\ v \in \hat{C},$$
$$x(\delta^+(v)) \leq 1 \quad \text{for } v \in \dot{C}.$$

Their sum gives

$$2 \sum_{a \in A(C)} x(a) - 2 \sum_{v \in \hat{C}} y(v) + \sum_{v \in \dot{C}} x(\delta^+(v) \setminus A(C)) + \sum_{v \in \tilde{C}} x(\delta^+(v) \setminus A(C)) \leq |A(C)| - 2|\hat{C}| + |\dot{C}|,$$

which implies

$$2 \sum_{a \in A(C)} x(a) - 2 \sum_{v \in \hat{C}} y(v) \leq |\tilde{C}| + |\dot{C}|.$$

Dividing by 2 and rounding down the right-hand side, we obtain

$$\sum_{a \in A(C)} x(a) - \sum_{v \in \hat{C}} y(v) \leq \frac{|\tilde{C}| + |\dot{C}| - 1}{2}. \qquad \square$$

Now we can present our main result.

THEOREM 7. *Let G be a directed graph; then $LP(G) = P(G)$ if and only if G does not contain an odd cycle.*

Proof. If G contains and odd cycle C, then we can define a vector $(\bar{x}, \bar{y}) \in P(G)$ as in (17)–(21). We have

$$\sum_{a \in A(C)} \bar{x}(a) - \sum_{v \in \hat{C}} \bar{y}(v) = \frac{|\tilde{C}| + |\hat{C}|}{2}.$$

Lemma 12 shows that $\bar{z} \notin LP(G)$.

Then the theorem follows from Theorem 4. □

6. Separation of odd cycle inequalities. Now we study the separation problem: Given a vector $(\bar{x}, \bar{y}) \in P(G)$, find an odd cycle inequality (22), if there is any, that separates $(\bar{x}, \bar{y})$ from $LP(G)$. These inequalities are $\{0, 1/2\}$-Chvátal–Gomory cuts, using the terminology of [4]. A separation algorithm can be obtained from the results of [4]. Here we give an alternative algorithm.

To solve the separation problem we write the inequalities as

$$2 \sum_{a \in A(C)} x(a) + \sum_{v \in \hat{C}} (1 - 2y(v)) \leq |A(C)| - 1$$

or

$$\sum_{a \in A(C)} (1 - 2x(a)) + \sum_{v \in \hat{C}} (2y(v) - 1) \geq 1.$$

In order to reduce this to a shortest path problem, several graph transformations are required.

6.1. First transformation. We build an auxiliary undirected graph $H = (N, F)$. For every arc $a = (u, v) \in A$ we create the nodes (u, a) and (v, a) in H. The first node is called a *tail* node, and the second is called a *head* node. The tail node is associated with u, and the head node is associated with v. We also create an edge between these two nodes with the weight $(1 - 2\bar{x}(u, v))$ and give the label *blue* to this edge; also this type of edge will be called *old*. See Figure 4.

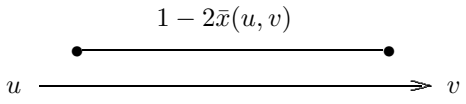


FIG. 4. *Edge associated with the arc (u, v). It has the label blue and is called old.*

Now for every node $v \in V$ and every pair of nodes in H associated with v we create an edge in H as follows. This type of edge will be called *new*. Let n_1 and n_2 be two nodes in H associated with v; we distinguish two cases:

- At least one of them is a tail node. In this case we add an edge between them with weight zero and label it *black*.
- Both n_1 and n_2 are head nodes. In this case we add an edge between them with weight $2\bar{y}(v) - 1$ and label this edge blue. See Figure 5.

A cycle in H consisting of an alternating sequence of old and new edges is called an *alternating* cycle. The separation problem reduces to finding an alternating cycle in H with an odd number of blue edges and total weight less than one.

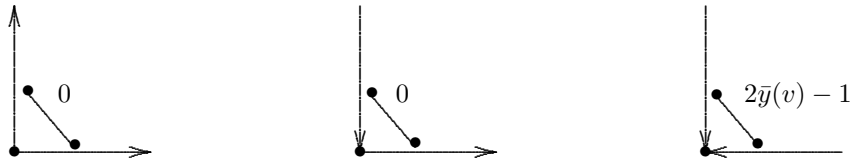


FIG. 5. *New edges. In the first two cases they have the label black, and in the last case it has the label blue. Beside each new edge we show their weight.*

6.2. Second transformation. To find an alternating cycle in H with an odd number of blue edges, we create a new graph $H' = (N', F')$ as follows. For every node $n \in H$ we make two copies n' and n''. Let n_1n_2 be an edge in H; we have two cases:

- If n_1n_2 is blue, we create the edges $n_1'n_2''$ and $n_1''n_2'$ with the same weight as n_1n_2 and the same name (old or new).
- If n_1n_2 is black, we create the edges $n_1'n_2'$ and $n_1''n_2''$ with the same weight as n_1n_2 and the same name (new).

Then for every node $n \in H$ we find a shortest alternating path P from n' to n'' in H'. The first edge in the path should be new, and the last edge should be old. Suppose that the weight of P is less than one; then for each node $p \in H$ such that p' and p'' are in P we identify them. This gives a (nonnecessarily simple) cycle that is alternating, has an odd number of blue edges, and has weight less than one. Notice that the derivation of inequalities (22) does not depend upon the cycle being simple.

Since the edge-weights could be negative, to find a shortest alternating path we have to modify the Bellman–Ford algorithm for shortest paths as follows. Let s be a source node. Let $f_o^k(v)$ be the length of a shortest alternating path from s to v having at most k arcs, whose first arc is new and whose last arc is old. Let $f_n^k(v)$ be the length of a shortest alternating path from s to v having at most k arcs, whose first arc is new and whose last arc is new. These values are computed with the following formulas:

$$
\begin{aligned}
&f_o^k(v) = \min\bigl\{f_o^{k-1}(v), \min\{f_n^{k-1}(u) + d_{uv} \mid uv \text{ is old}\}\bigr\},\\
&f_n^k(v) = \min\bigl\{f_n^{k-1}(v), \min\{f_o^{k-1}(u) + d_{uv} \mid uv \text{ is new}\}\bigr\},\\
&f_o^0(s) = 0, \quad f_n^0(s) = \infty,\\
&f_o^0(v) = f_n^0(v) = \infty \quad \text{for } v \neq s.
\end{aligned}
$$

This algorithm requires that the graph has no alternating cycle of negative weight; this is shown below.

LEMMA 13. *The edge weights cannot create a cycle of negative weight.*

Proof. Suppose that

$$\sum_{a \in A(C)} (1 - 2\bar{x}(a)) + \sum_{v \in \hat{C}} \bigl(2\bar{y}(v) - 1\bigr) < 0$$

for some cycle C. This implies

$$2 \sum_{a \in A(C)} \bar{x}(a) - 2 \sum_{v \in \hat{C}} \bar{y}(v) > |C| - |\hat{C}|,$$

but when deriving inequalities (22) we had

$$2 \sum_{a \in A(C)} \bar{x}(a) - 2 \sum_{v \in \hat{C}} \bar{y}(v) \leq |C| - |\hat{C}|. \qquad \square$$

The complexity of this method is as follows.

THEOREM 8. *The separation problem for inequalities* (22) *can be solved in* $O(|V|^2|A|^2)$ *time.*

Proof. For the graph $H = (N, F)$, we have $|N| = 2|A|$ and $|F| \leq |A| + |A||V|$. For $H' = (N', F')$, we have $|N'| = 4|A|$ and $|F'| \leq 2|A| + 2|A||V|$. For a particular value k, computing the values f takes $O(|F'|)$ operations. Since $k \leq |V|$, applying this algorithm for a particular source s takes $O(|V|^2|A|)$ operations. Since every node of H should be tried as a source, the entire procedure takes $O(|V|^2|A|^2)$ time. ∎

7. Detecting odd cycles. Now we study how to recognize the graphs G for which $LP(G) = P(G)$. We start with a graph G, and a new undirected graph $H = (N, E)$ is built as follows. For every node $u \in G$ we have the nodes u' and u'' in N and the edge $u'u'' \in E$. For every arc $(u, v) \in G$ we have an edge $u'v'' \in E$. See Figure 6.

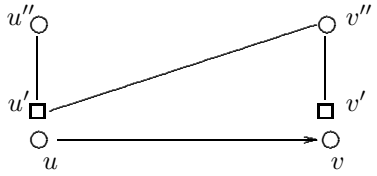


FIG. 6. *Basic transformation to create the graph* H.

Considering a cycle C in G, we build a cycle C_H in H as follows:

- If (u, v) and (u, w) are in C, then the edges $u'v''$ and $u'w''$ are taken.
- If (u, v) and (w, v) are in C, then the edges $u'v''$ and $v''w'$ are taken.
- If (u, v) and (v, w) are in C, then the edges $u'v''$, $v''v'$, and $v'w''$ are taken.

On the other hand, a cycle in H corresponds to a cycle in G. Thus there is a one to one correspondence among cycles of G and cycles of H. Moreover, if the cycle in H has cardinality $2q$, then $q = |\dot{C}| + |\tilde{C}|$, where C is the corresponding cycle in G. Therefore, an odd cycle in G corresponds to a cycle in H of cardinality $2(2p + 1)$ for some positive integer p. See Figure 7.

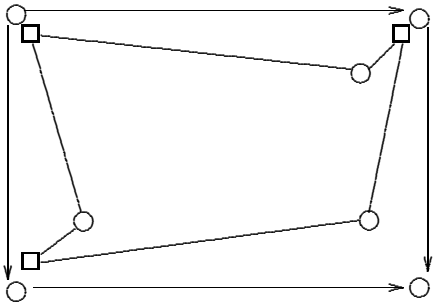


FIG. 7. *An odd cycle in* G *and the corresponding cycle in* H. *The nodes of* H *close to a node* $u \in G$ *correspond to* u' *or* u''.

In other words, finding an odd cycle in G reduces to finding a cycle of cardinality $2(2p + 1)$ for some positive integer p in the bipartite graph H.

For this question, a linear time algorithm was given in [19]. A simple $O(|V||A|^2)$ has been given in [9]; we describe it below.

First we should find a cycle basis of H and test if the cardinality of every cycle in this basis is 0 mod 4. If there is one whose cardinality is 2 mod 4, we are done. Otherwise consider the symmetric difference of two cycles whose cardinality is 0 mod 4. If the cardinality of their intersection is even, then the cardinality of their symmetric difference is 0 mod 4; otherwise it is 2 mod 4. Since any cycle C can be obtained as the symmetric difference of a set of cycles in the basis, if the cardinality of C is 2 mod 4, then there are at least two cycles in the basis whose symmetric difference has cardinality 2 mod 4. Therefore, one just has to test all elements of a cycle basis and the symmetric difference of all pairs.

8. Uncapacitated facility location. Now we assume that V is partitioned into V_1 and V_2, $A \subseteq V_1 \times V_2$, and we deal with the system

$$\sum_{(u,v)\in A} x(u,v) = 1 \quad \forall u \in V_1, \tag{23}$$

$$x(u,v) \le y(v) \quad \forall (u,v) \in A, \tag{24}$$

$$0 \le y(v) \le 1 \quad \forall v \in V_2, \tag{25}$$

$$x(u,v) \ge 0 \quad \forall (u,v) \in A. \tag{26}$$

We denote by $\Pi(G)$ the polytope defined by (23)–(26). Notice that $\Pi(G)$ is a face of $P(G)$. Let $\bar{V}_1$ be the set of nodes $u \in V_1$ with $|\delta^+(u)| = 1$. Let $\bar{V}_2$ be the set of nodes in V_2 that are adjacent to a node in $\bar{V}_1$. It is clear that the variables associated with nodes in $\bar{V}_2$ should be fixed, i.e., $y(v) = 1$ for all $v \in \bar{V}_2$. Let us denote by $\bar{G}$ the subgraph induced by $V \setminus \bar{V}_2$. In this section we prove that $\Pi(G)$ is an integral polytope if and only if $\bar{G}$ has no odd cycle.

Let us first assume that $\bar{G}$ has no odd cycle. As before, we suppose that $\bar{z}$ is a fractional extreme point of $\Pi(G)$. The analogues of Lemmas 1–4 apply here. Thus we can assume that we deal with a connected component G'. Lemma 2 implies that any node in $\bar{V}_2$ is not in a cycle of G'. Therefore, G' has no odd cycle and $P(G')$ is an integral polytope. Since $\Pi(G')$ is a face of $P(G')$, we have a contradiction.

Now let C be an odd cycle of $\bar{G}$. We can define a fractional vector as follows:

$$\begin{aligned}
&\bar{y}(v) = 1/2 \quad \forall \text{ nodes } v \in V_2 \cap V(C),\\
&\bar{x}(a) = 1/2 \quad \text{for } a \in A(C),\\
&\bar{y}(v) = 1 \quad \forall \text{ nodes } v \in V_2 \setminus V(C).
\end{aligned}$$

For every node $u \in V_1 \setminus V(C)$, we look for an arc $(u,v) \in \delta^+(u)$. If $\bar{y}(v) = 1$, we set $\bar{x}(u,v) = 1$. If $\bar{y}(v) = 1/2$, then there is another arc $(u,w) \in \delta^+(u)$ such that $\bar{y}(w) = 1/2$ or $\bar{y}(w) = 1$. We set $\bar{x}(u,v) = \bar{x}(u,w) = 1/2$. Finally, we set $\bar{x}(a) = 0$ for each remaining arc a. This vector satisfies (23)–(26), but it violates the inequality (22) associated with C. So in this case (23)–(26) does not define an integral polytope. Thus we can state the following.

THEOREM 9. *The system* (23)–(26) *defines an integral polytope if and only if* $\bar{G}$ *has no odd cycle.*

THEOREM 10. *The UFLP is polynomially solvable for graphs* G *such that* $\bar{G}$ *has no odd cycle.*

This class of bipartite graphs can be recognized in polynomial time as described in section 7.

Acknowledgments. We are grateful to Gérard Cornuéjols for pointing out references [19] and [9]. We also thank the referees for their helpful comments.

REFERENCES

[1] M. Baïou and F. Barahona, *On the p-median polytope of Y-free graphs*, Discrete Optim., 5 (2008), pp. 205–219.
[2] J. Byrka and K. Aardal, *The approximation gap for the metric facility location problem is not yet closed*, Oper. Res. Lett., 35 (2007), pp. 379–384.
[3] L. Cánovas, M. Landete, and A. Marín, *On the facets of the simple plant location packing polytope*, Discrete Appl. Math., 124 (2002), pp. 27–53.
[4] A. Caprara and M. Fischetti, $\{0, \frac{1}{2}\}$-*Chvátal-Gomory cuts*, Math. Programming, 74 (1996), pp. 221–235.
[5] D. C. Cho, E. L. Johnson, M. Padberg, and M. R. Rao, *On the uncapacitated plant location problem.* I. *Valid inequalities and facets*, Math. Oper. Res., 8 (1983), pp. 579–589.
[6] D. C. Cho, M. W. Padberg, and M. R. Rao, *On the uncapacitated plant location problem.* II. *Facets and lifting theorems*, Math. Oper. Res., 8 (1983), pp. 590–612.
[7] F. A. Chudak and D. B. Shmoys, *Improved approximation algorithms for the uncapacitated facility location problem*, SIAM J. Comput., 33 (2003), pp. 1–25.
[8] V. Chvátal, *On certain polytopes associated with graphs*, J. Combin. Theory Ser. B, 18 (1975), pp. 138–154.
[9] M. Conforti and M. R. Rao, *Structural properties and recognition of restricted and strongly unimodular matrices*, Math. Programming, 38 (1987), pp. 17–27.
[10] G. Cornuejols, M. L. Fisher, and G. L. Nemhauser, *Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms*, Management Sci., 23 (1976/1977), pp. 789–810.
[11] G. Cornuejols and J.-M. Thizy, *Some facets of the simple plant location polytope*, Math. Programming, 23 (1982), pp. 50–74.
[12] C. De Simone and C. Mannino, *Easy Instances of the Plant Location Problem*, Technical report R. 427, IASI-CNR, Rome, Italy, 1996.
[13] M. Guignard, *Fractional vertices, cuts and facets of the simple plant location problem*, Math. Programming Stud., 12 (1980), pp. 150–162.
[14] P. B. Mirchandani and R. L. Francis, eds., *Discrete Location Theory*, Wiley-Intersci. Ser. Discrete Math. Optim., Wiley, New York, 1990.
[15] R. Müller and A. S. Schulz, *Transitive packing: A unifying concept in combinatorial optimization*, SIAM J. Optim., 13 (2002), pp. 335–367.
[16] D. B. Shmoys, É. Tardos, and K. Aardal, *Approximation algorithms for facility location problems (extended abstract)*, in Proceedings of the 29th ACM Symposium on Theory of Computing, ACM, New York, 1997, pp. 265–274.
[17] M. Sviridenko, *An improved approximation algorithm for the metric uncapacitated facility location problem*, in Integer Programming and Combinatorial Optimization, Lecture Notes in Comput. Sci. 2337, Springer-Verlag, Berlin, 2002, pp. 240–257.
[18] J. Vygen, *Approximation Algorithms for Facility Location Problems*, Technical report, University of Bonn, Bonn, Germany, 2005.
[19] M. Yannakakis, *On a class of totally unimodular matrices*, Math. Oper. Res., 10 (1985), pp. 280–304.

SIAM J. DISCRETE MATH.
Vol. 23, No. 2, pp. 680–693

STEINER TREES AND CONVEX GEOMETRIES*

MORTEN H. NIELSEN† AND ORTRUD R. OELLERMANN†

Abstract. Let V be a finite set and $\mathcal{M}$ a collection of subsets of V. Then $\mathcal{M}$ is an alignment of V if and only if $\mathcal{M}$ is closed under taking intersections and contains both V and the empty set. If $\mathcal{M}$ is an alignment of V, then the elements of $\mathcal{M}$ are called convex sets and the pair $(V, \mathcal{M})$ is called an aligned space. If $S \subseteq V$, then the convex hull of S is the smallest convex set that contains S. Suppose $X \in \mathcal{M}$. Then $x \in X$ is an extreme point for X if $X \setminus \{x\} \in \mathcal{M}$. The collection of all extreme points of X is denoted by $ex(X)$. A convex geometry on a finite set is an aligned space with the additional property that every convex set is the convex hull of its extreme points. Let G be a connected graph. A set S of vertices is g-convex if for every pair u, v of vertices in S, every vertex that belongs to some u-v geodesic (shortest path) is also in S. A set S of vertices in G is k-Steiner-convex, denoted by g_k-convex, if, for every set T of k vertices of S, every vertex that belongs to some Steiner tree for T, i.e., a subtree of G of smallest size containing T, is also in S. Let $R = \{k_1, k_2, \ldots, k_t\}$ be a collection of positive integers such that $2 \leq k_1 < k_2 < \cdots < k_t$. We say a set S of vertices in a connected graph is g_R-convex if S is g_{k_i}-convex for $1 \leq i \leq t$. A set S of vertices of G is g^3-convex if, for every pair u, v of vertices of S, distance at least 3 apart in G, every vertex that belongs to some u-v geodesic in G is also in S. A set of vertices that is both g^3-convex and g_3-convex is called a g_3^3-convex set. Structural characterizations are given of those classes of graphs for which (i) the g_3-convex sets, (ii) the g_R-convex sets for those sets R that have minimum element 2 or 3, and (iii) the g_3^3-convex sets form a convex geometry.

Key words. Steiner distance, Steiner intervals, Steiner convex sets, convex geometries

AMS subject classifications. 05C75, 05C12, 05C17

DOI. 10.1137/070691383

1. Introduction.

This paper is motivated by the results and ideas contained in [7, 8]. We introduce new graph convexities and show how these give rise to structural characterizations of certain graph classes. For graph terminology we follow [3] and [5]. All graphs considered here are connected, finite, simple (i.e., without loops and multiple edges), unweighted, and undirected. The structural characterizations of graphs that we describe are often given in terms of forbidden subgraphs. Let G and F be graphs. Then F is an *induced subgraph* of G if F is a subgraph of G and for every $u, v \in V(F)$, $uv \in E(F)$ if $uv \in E(G)$. We say a graph G is *F-free* if it does not contain F as an induced subgraph. Suppose $\mathcal{C}$ is a collection of graphs. Then G is $\mathcal{C}$-free if G is F-free for every $F \in \mathcal{C}$. If F is a path or cycle that is a subgraph of G, then F has a *chord* if it is not an induced subgraph of G; i.e., F has two vertices that are adjacent in G but not in F. An induced cycle of length at least 5 is called a *hole*.

When it is clear from context which graph is being considered, we denote by $N(v)$ the set of neighbors (i.e., the *neighborhood*) of a given vertex v in the graph. Further, we use $N[v]$ to denote the *closed neighborhood* of the vertex v, i.e., the set $N(v) \cup \{v\}$. If S is a subgraph of G or a subset of $V(G)$, then $N_S(v)$ denotes the set of neighbors of v in S.

We begin with an overview of convexity notions in graphs. For a more extensive overview of other abstract convex structures, see [16].

*Received by the editors May 11, 2007; accepted for publication (in revised form) November 24, 2008; published electronically March 4, 2009.

http://www.siam.org/journals/sidma/23-2/69138.html

†University of Winnipeg, 515 Portage Avenue, Winnipeg, MB R3B 2E9, Canada (m.nielsen@uwinnipeg.ca, o.oellermann@uwinnipeg.ca). The second author was supported by an NSERC grant Canada.

Let V be a finite set and $\mathcal{M}$ a collection of subsets of V. Then $\mathcal{M}$ is an *alignment* of V if and only if $\mathcal{M}$ is closed under taking intersections and contains both V and the empty set. If $\mathcal{M}$ is an alignment of V, then the elements of $\mathcal{M}$ are called *convex sets* and the pair $(V, \mathcal{M})$ is called an *aligned space*. If $S \subseteq V$, then the *convex hull* of S is the smallest convex set that contains S. Suppose $X \in \mathcal{M}$. Then $x \in X$ is an *extreme point* for X if $X \setminus \{x\} \in \mathcal{M}$. The collection of all extreme points of X is denoted by $ex(X)$. A *convex geometry* on a finite set V is an aligned space $(V, \mathcal{M})$ with the additional property that every convex set is the convex hull of its extreme points. This property is referred to as the *Minkowski–Krein–Milman (MKM)* property.

Farber and Jamison [8] established the following fundamental result for convex geometries.

THEOREM 1. *Suppose $(V, \mathcal{M})$ is a convex geometry. Then $S \in \mathcal{M}$ if and only if there exists an ordering $(v_1, v_2, \ldots, v_k)$ of $V \setminus S$ such that v_i is an extreme point of $S \cup \{v_i, v_{i+1}, \ldots, v_k\}$ for each $i = 1, 2, \ldots, k$.*

For a given ordering $(v_1, v_2, \ldots, v_n)$ of the vertex set V of a graph G, let $G_i = \langle\{v_i, v_{i+1}, \ldots, v_n\}\rangle$; i.e., G_i is the subgraph induced by $\{v_i, v_{i+1}, \ldots, v_n\}$. Several classes of graphs can be characterized in terms of vertex orderings as follows: A graph G belongs to a class $\mathcal{G}$ if and only if there is an ordering $(v_1, v_2, \ldots, v_n)$ of $V(G)$ such that v_i has property $\mathbf{P}$ in G_i for $i = 1, 2, \ldots, n$. In that case we say that the ordering $(v_1, v_2, \ldots, v_n)$ is a $\mathbf{P}$ *elimination ordering* for G or simply a $\mathbf{P}$ *ordering* for G. For example, if $\mathbf{P}$ is the property "has a complete neighborhood," then $\mathcal{G}$ is the class of chordal graphs (see [3]). Theorem 1 suggests that such classes of graphs may be related to convex geometries. In particular, we will be interested in properties $\mathbf{P}$ that describe the extreme vertices with respect to a given graph convexity. Moreover, for a given collection $\mathcal{M}$ of subsets of the vertex set of a graph G, we are interested in determining when $(V(G), \mathcal{M})$ is a convex geometry.

Several abstract convexities associated with the vertex set of a graph are well known (see [8]). Their study is of interest in computational geometry and has some direct applications to other areas—for example, game theory (see [2]). For another text containing material on graph convexity, see [3].

We next discuss graph convexities whose convex sets are described in terms of induced paths (i.e., paths without chords) having certain properties. The *distance* between a pair of vertices u, v of G is the length of a shortest u-v path in G and is denoted by $d_G(u, v)$ or, if G is clear from context, simply $d(u, v)$. The *eccentricity* $ecc(v)$ of a vertex v in G is the maximum distance between v and any other vertex in G. A vertex at distance $ecc(v)$ from v is said to be an *eccentric vertex for* v. A shortest u-v path is also called a u-v *geodesic*. Geodesics are necessarily induced paths, but not all induced paths are geodesics. The *g-interval* (respectively, *m-interval*) between a pair u, v of vertices in a graph G is the collection of all vertices that lie on some u-v geodesic (respectively, induced u-v path) in G and is denoted by $I_g^{(G)}[u, v]$ (respectively, $I_m^{(G)}[u, v]$). When G is clear from context, the superscript (G) will be omitted.

A subset S of vertices of a graph is said to be *g-convex* (*m-convex*) if it contains the g-interval (m-interval) between every pair of vertices in S. It is not difficult to see that the collection of all g-convex (m-convex) sets is an alignment of V. A vertex in a graph is *simplicial* if its neighborhood induces a complete subgraph. It is well known that a graph G has a *simplicial elimination ordering* (also called a *perfect elimination ordering*) if and only if it is chordal, i.e., G has no induced cycles of length more than 3. It can readily be seen that v is an extreme point for a g-convex or m-convex

set S if and only if v is simplicial in the subgraph induced by S. Of course, the convex hull of the extreme points of a convex set S is contained in S, but equality holds only in special cases. In [8] those graphs for which the g-convex sets form a convex geometry are characterized.

THEOREM 2. *Let $G = (V, E)$ be a connected graph and let $\mathcal{M}_g(G)$ be the collection of g-convex sets of G. Then $(V, \mathcal{M}_g(G))$ is a convex geometry if and only if G is chordal and has no induced 3-fan (see Figure 1).*

Chordal graphs without induced 3-fans are also known as the ptolemaic graphs and are precisely the chordal, distance-hereditary graphs. (A graph is *distance-hereditary* if for every connected induced subgraph H of G and every pair u, v of vertices of H, $d_H(u, v) = d_G(u, v)$.) Moreover, in [8] those graphs for which the m-convex sets form a convex geometry are characterized as precisely the chordal graphs.

For what follows we use P_k to denote an induced path of order k. A vertex is simplicial in a set S of vertices if and only if it is not the central vertex of a P_3 in $\langle S \rangle$. Jamison and Olariu [10] relaxed this condition: They defined a vertex to be *semisimplicial* in S if and only if it is not a central vertex of a P_4 in $\langle S \rangle$.

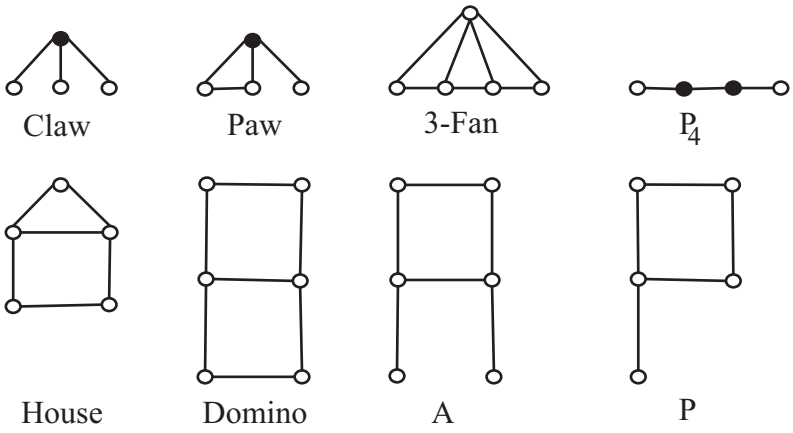


FIG. 1. *Some special graphs.*

Dragan, Nicolai, and Brandstädt in [7] introduced another convexity notion that relies on induced paths. The m^3-*interval* between a pair u, v of vertices in a graph G, denoted by $I_{m^3}[u, v]$, is the collection of all vertices of G that belong to an induced u-v path of length at least 3. Let G be a graph with vertex set V. A set $S \subseteq V$ is m^3-*convex* if and only if, for every pair u, v of vertices of S, the vertices of the m^3-interval between u and v belong to S. It is not difficult to see that the collection of all m^3-convex sets is an alignment. Note that an m^3-convex set is not necessarily connected. It is shown that the extreme points of an m^3-convex set are precisely the semisimplicial vertices of $\langle S \rangle$. Moreover, those graphs for which the m^3-convex sets form a convex geometry are characterized in [7].

THEOREM 3. *Let $G = (V, E)$ be a connected graph and let $\mathcal{M}_{m^3}(G)$ be the collection of m^3-convex sets of G. Then the following are equivalent:*

(1) *G is (house, hole, domino, A)-free.*

(2) *$(V, \mathcal{M}_{m^3}(G))$ is a convex geometry.*

If G is a graph of order n, there are $n!$ orderings of its vertices. It is thus not clear, for a given property $\mathbf{P}$, that there is an efficient procedure for recognizing if a graph has a $\mathbf{P}$ elimination ordering. Several linear-time search techniques have been proposed, two of which we describe here. Rose, Tarjan, and Leuker [14] proposed the first of these, namely, the *lexicographic breadth-first-search (LexBFS).*

LexBFS. Order the vertices of a graph G by assigning them numbers from $|V|$ to 1 as follows: For k from $n = |V|$ down to 1, assign the number k to an as yet unnumbered vertex v which has a lexicographically largest vector $(s_n, s_{n-1}, \ldots, s_{k+1})$, where $s_i = 1$ if v is adjacent to a vertex numbered i and $s_i = 0$ otherwise for $k+1 \leq i \leq n$. It is assumed that initially every vector is empty. So LexBFS may begin at any vertex.

The second search technique we describe is due to Tarjan and Yannakakis [15] and is called the *maximum cardinality search (MCS).*

MCS. Order the vertices of a graph G by assigning them numbers from $|V|$ to 1 as follows: For k from $n = |V|$ down to 1, assign the number k to an as yet unnumbered vertex that is adjacent to a maximum number of numbered vertices.

Jamison and Olariu showed in [10] that the graphs for which every LexBFS ordering is a semisimplicial ordering are precisely the *HHD*-free graphs, i.e., the (house, hole, domino)-free graphs. Moreover, they characterized the graphs for which every MCS ordering of every induced subgraph is a semisimplicial ordering as the *HHP*-free graphs, i.e., the (house, hole, P)-free graphs.

A set S of vertices in a graph G is g^3*-convex* if, for every pair $u, v \in S$ such that $d_G(u, v) \geq 3$, $I[u, v] \subseteq S$. A vertex v in a graph G is defined in [12] to be *weakly semisimplicial* in S if and only if, for all $u, w \in N_S(v)$, one of the following three conditions holds:

(1) $uw \in E(G)$,
(2) $uw \notin E(G)$ and $N_S(u) \setminus N_S(v) = N_S(w) \setminus N_S(v)$,
(3) $uw \notin E(G)$ and $N_S(u) \setminus N_S(v) \neq N_S(w) \setminus N_S(v)$, and for every x in $N_S(w) \setminus (N_S(v) \cup N_S(u))$ we have $N_G(x) \cap N_G(u) \neq \emptyset$, and for every y in $N_S(u) \setminus (N_S(v) \cup N_S(w))$ we have $N_G(y) \cap N_G(w) \neq \emptyset$.

A vertex satisfying condition (1) alone is simplicial. So condition (1) characterizes extreme points of g- and m-convex sets. A vertex satisfying condition (1) or (2) is semisimplicial. So these two conditions characterize extreme vertices of m^3-convex sets. It is shown in [12] that v is an extreme point of a g^3-convex set if and only if v is weakly semisimplicial. Note that every semisimplicial vertex is weakly semisimplicial.

We now introduce a graph convexity that generalizes g-convexity. The *Steiner interval* of a set S of vertices in a connected graph G, denoted by $I(S)$, is the union of all vertices of G that lie on some *Steiner tree* for S, i.e., a connected subgraph that contains S and has the minimum number of edges among all such subgraphs. Steiner intervals have been studied, for example, in [11, 13]. A set S of vertices in a graph G is *k-Steiner-convex*, denoted by g_k-convex, if the Steiner interval of every collection of k vertices of S is contained in S. Thus S is g_2-convex if and only if it is g-convex. The collection of g_k-convex sets forms an aligned space. We call an extreme point of a g_k-convex set a *k-Steiner simplicial* vertex, abbreviated kSS vertex.

The extreme points of g_3-convex sets S, i.e., the $3SS$ vertices, are characterized in [4] as those vertices that are *not* a central vertex of an induced claw, paw, or P_4 in $\langle S \rangle$ (see Figure 1). Thus a $3SS$ vertex is semisimplicial and hence weakly semisimplicial. In [4] those graphs for which every LexBFS ordering is a $3SS$ ordering and those for which every MCS ordering of every induced subgraph is a $3SS$ ordering are characterized.

Some of the previous convexity notions may be combined in a natural way to

obtain new convexity notions for graphs. Suppose $R = \{k_1, k_2, \ldots, k_t\}$ is a collection of positive integers such that $2 \leq k_1 < k_2 < \cdots < k_t$. We say a set S of vertices in a connected graph is g_R-convex if S is g_{k_i}-convex for $1 \leq i \leq t$. It is readily seen that the collection of g_R-convex sets forms an alignment of $V(G)$. Moreover, v is an extreme vertex of a g_R-convex set S if and only if v is k_iSS in S for every $k_i \in R$. Since simplicial vertices are kSS for every $k \geq 3$, it follows that if $2 \in R$, then the extreme points of a g_R-convex set are precisely the simplicial vertices of the set.

We propose a further convexity notion that combines g_3-convexity and g^3-convexity. A graph is g_3^3-convex if it is both g^3- and g_3-convex. Since the $3SS$ vertices of a g_3^3-convex set are weakly semisimplicial, the extreme points of a g_3^3-convex set are precisely the $3SS$ vertices. We give structural characterizations of those graphs G for which $(V(G), \mathcal{M})$ is a convex geometry, where $\mathcal{M}$ is (i) the collection of all g_3-convex sets; (ii) the collection of all g_R-convex sets where R is a set of positive integers, each at least 2, and where $2 \in R$ or $3 \in R$; and (iii) the collection of all g_3^3-convex sets.

The following results will be useful in what follows. We begin with a structural characterization of distance-hereditary graphs given in [9]. Suppose C is a cycle and e and f are two chords of C. If $C + e + f$ is homeomorphic to K_4, then we say e and f are *crossing chords*.

THEOREM 4. *A graph G is distance-hereditary if and only if every cycle of length at least 5 in G has a pair of crossing chords.*

Another useful characterization of distance-hereditary graphs is given in [1].

THEOREM 5. *A connected graph G is distance-hereditary if and only if it is (house, hole, domino, 3-fan)-free.*

As an immediate consequence we have the following.

COROLLARY 6. *Let G be a graph with $diam(G) \leq 2$. Then G is distance-hereditary if and only if G is (house, hole, 3-fan)-free.*

COROLLARY 7. *If G is (house, hole, domino, 3-fan)-free, then a set S of vertices is g^3-convex if and only if it is m^3-convex.*

Proof. If S is m^3-convex, then S is g^3-convex. Also, by Theorem 5, if S is g^3-convex, then S contains all induced paths between pairs of vertices of S of length at least 3, since such paths are geodesics. Thus S is m^3-convex. □

Let $k \geq 2$ be an integer. A graph G is defined to be k-Steiner distance-hereditary if, for every connected induced subgraph H of G and every set S of k vertices in H, the Steiner distance of S in H is the same as the Steiner distance of S in G. The following result was established in [6].

THEOREM 8. *If G is distance-hereditary, then G is k-Steiner distance-hereditary for every integer $k \geq 2$.*

2. Convex geometries.

2.1. g_3-convex geometries. Let $G = (V, E)$ be a connected graph and let $\mathcal{M}_{g_3}(G)$ be the collection of g_3-convex sets. In this section we determine the class of connected graphs G for which $(V, \mathcal{M}_{g_3}(G))$ is a convex geometry.

A graph G is a *replicated-twin C_4* if it is isomorphic to any one of the four graphs shown in Figure 3(a), where any subset of the dotted edges may belong to G. The collection of the four replicated-twin C_4 graphs is denoted by $\mathcal{R}_{C_4}$. For a set S of vertices in a graph G, the g_3-convex hull of S is denoted by $g_3\text{-}conv(S)$.

THEOREM 9. *Let $G = (V, E)$ be a graph. Then the following are equivalent:*

(1) *G is (P_4, $\mathcal{R}_{C_4}$)-free.*

(2) *$(V, \mathcal{M}_{g_3}(G))$ is a convex geometry.*

Proof. Suppose $(V, \mathcal{M}_{g_3}(G))$ is a convex geometry. Suppose first that G contains an induced P_4, say, $P = uv_1v_2v$. Let S be the g_3-convex hull of $V(P)$. Since u and v are the only $3SS$ vertices of P, the $3SS$ vertices of S are a subset of $\{u, v\}$. But the g_3-convex hull of any subset T of $\{u, v\}$ is just T and hence does not contain all the vertices of S, contradicting the fact that $(V, \mathcal{M}_{g_3}(G))$ is a convex geometry. Suppose next that G contains a replicated-twin C_4, say, H, as an induced subgraph. Then H contains no $3SS$ vertices. Let $S = g_3\text{-conv}(V(H))$. Since H has no $3SS$ vertices, S has no $3SS$ vertices, and S is therefore not the g_3-convex hull of its $3SS$ vertices, again contradicting the fact that $(V, \mathcal{M}_{g_3}(G))$ is a convex geometry. Thus (2) implies (1).

Suppose there is a graph G for which (1) but not (2) holds. Since G is P_4-free, $diam(G) \leq 2$ and G is (house, hole, domino, 3-fan)-free. Let G be such a graph of smallest possible order. We may assume that $diam(G) = 2$, for if $diam(G) = 1$, then G is complete and (2) holds. Then any proper connected induced subgraph of G has the property that it is the g_3-convex hull of its extreme points. (Note that, by Theorem 5, G is distance-hereditary, so this subgraph also has diameter at most 2.) One can check, if G has order at most 4, that $(V, \mathcal{M}_{g_3}(G))$ is a convex geometry. Suppose thus that $|V| \geq 5$.

Case 1. The radius of G is 1. Then G has a vertex v that is adjacent to every other vertex of G.

Subcase 1.1. $G - v$ is connected. By our choice of G, the g_3-convex hull of the extreme vertices of $G - v$ is $V(G - v)$. Suppose w is an extreme point of $V(G - v)$. Then w is not the central vertex of an induced claw or paw in $G-v$. Since v is adjacent to every other vertex of G, it is not a peripheral vertex of any induced claw or paw. Hence w is an extreme point of G also. So the collection of extreme points of G, $ex(V)$, contains the collection of extreme points of $G - v$, i.e., $ex(V(G-v))$. By Corollary 6, G is distance-hereditary, and thus, by Theorem 8, G is Steiner distance-hereditary. Thus $I_{G-v}(S) \subseteq I_G(S)$ for every set S of vertices in $G - v$. This holds in particular if S is a set of three vertices of $G - v$. Therefore, $V(G - v) = g_3\text{-conv}(ex(V(G - v))) \subseteq g_3\text{-conv}(ex(V))$. So if v is an extreme point of G, $V = g_3\text{-conv}(ex(V))$. If v is not an extreme point of G, it is the central vertex of a claw or paw whose three peripheral vertices are contained in $G - v$. So v is in the Steiner interval of these three vertices and thus in the g_3-convex hull of the extreme vertices of G.

Subcase 1.2. $G-v$ is disconnected. Let $H_1, H_2, \ldots, H_k$, $k \geq 2$, be the components of $G - v$. Since G is distance-hereditary, $diam(H_i) \leq 2$. By the choice of G, each $V(H_i)$ is the g_3-convex hull of its extreme points. As in Subcase 1.1, the extreme points of H_i, $1 \leq i \leq k$, are contained in the extreme points of G. By our choice of G, the g_3-convex hull (in G) of $ex(V(H_i))$ is either $V(H_i)$ or $V(H_i) \cup \{v\}$ (for $1 \leq i \leq k$). If, for some i with $1 \leq i \leq k$, H_i has at least two vertices, then H_i contains at least two extreme points, say, x and y. Let z be an extreme point of H_j for some $j \neq i$. Then v is in the Steiner interval for $\{x, y, z\}$. Hence v is in $g_3\text{-conv}(ex(V))$. We may thus assume that each H_i contains exactly one vertex. Since G has at least five vertices, this implies that $G-v$ has at least four components. Thus v is in the Steiner interval of three extreme points of G chosen from distinct components of $G-v$. Thus again $g_3\text{-conv}(ex(V)) = V$.

Case 2. The radius of G is 2. Let u be any vertex in G. Since each vertex has eccentricity 2, there exists a vertex u' such that $d(u, u') = 2$. Let $S = N(u) \cap N(u')$. If u has a neighbor w such that $d(w, u') = 2$, then $S \subseteq N(w)$; otherwise, if $r \in S \setminus N(w)$, then $wuru'$ is an induced P_4 (which is not possible in a distance-hereditary graph of diameter two). Similarly, if u' has a neighbor w such that $d(w, u) = 2$, then $S \subseteq N(w)$.

By our choice of G, there exists a vertex $v \in V$ which is not $3SS$; i.e., v is the

central vertex of an induced paw or claw $\langle\{v, x, y, z\}\rangle$ for some vertices x, y, z. Among all vertices u that are the central vertex of some paw or claw induced by $\{u, x, y, z\}$, let v be one of maximum degree in G. We may assume that $zx, zy \notin E$. Let v' be an eccentric vertex for v.

Let $S = N(v) \cap N(v')$. By the above observation, either $\{x, y, z\} \subseteq S$ or $d(x, v') = d(y, v') = d(z, v') = 2$.

Subcase 2.1. $\{x, y, z\} \subseteq S$. We show first that $N(v) = N(v')$. Suppose $w \in N(v) \setminus N(v')$. Then w is adjacent with each of x, y, and z; otherwise, G contains an induced P_4, $wvuv'$, for some $u \in \{x, y, z\}$. But now $\{v, w, x, y, z, v'\}$ induces a replicated-twin C_4, which is a contradiction. Hence $N(v) \subseteq N(v')$ and, by symmetry, $N(v') \subseteq N(v)$.

We now show that $V \setminus \{v, v'\} = S$. Suppose $r \in V \setminus (S \cup \{v, v'\})$. If $ru \in E$ for $u = x$ or y, then $rz \in E$ (otherwise, $ruvz$ is an induced P_4). If $rz \in E$, then $ru \in E$ for $u = x$ and y (otherwise, $rzvu$ is an induced P_4 for $u = x$ or y). Hence if $ru \in E$ for some $u \in \{x, y, z\}$, then $\{x, y, z\} \subseteq N(r)$; but then $\langle\{v, r, v', x, y, z\}\rangle$ is a replicated-twin C_4. So r must be nonadjacent with each of x, y, and z.

Since $d(r, v') = 2$, there exists $w \in N(r) \cap (S \setminus \{x, y, z\})$ and $wu \in E$ for all $u = x, y, z$ (otherwise, $\langle\{r, w, v, u\}\rangle$ is a P_4). Now, as w is the central vertex of the claw or paw $\langle\{w, x, y, z\}\rangle$ and $\deg w \geq |\{r, v, x, y, z, v'\}| = 6$, by the choice of v, there must exist two vertices w' and w'' in $S \setminus N[w]$. Then $rw', rw'' \in E$ (otherwise, $w'vwr$ or $w''vwr$ is an induced P_4); however, this implies that $\langle\{r, v, v', w, w', w''\}\rangle$ is a replicated-twin C_4. Hence $V \setminus \{v, v'\} = S$.

In particular, $G - v$ is connected. Suppose every extreme point of $V(G - v)$ is also an extreme point of V. By our choice of G, the g_3-convex hull of the extreme points (i.e., the $3SS$ vertices) of $G - v$ is $V(G - v)$ and thus contains x, y, and z. By Theorem 5, this implies that x, y, z also belong to the g_3-convex hull of the extreme points of V, and hence so does v, which is a contradiction. Therefore, there is some extreme point u of $V(G - v)$ that is not an extreme point of V. Thus u is the central vertex of some paw or claw in G but not in $G - v$; i.e., this paw or claw contains v. Let $\{u, v, s, t\}$ be the vertices of this paw or claw. Then, since either s or t must belong to S, $\langle\{s, t, v\}\rangle$ contains at least two edges, which contradicts the assumption that $\langle\{u, v, s, t\}\rangle$ is a claw or paw.

Subcase 2.2. $d(x, v') = d(y, v') = d(z, v') = 2$. Let $u \in S$. Then u is adjacent with every $t \in \{x, y, z\}$ (otherwise, $tvuv'$ is an induced P_4), so u must have degree at least $|S| - 2$ in $\langle S\rangle$ (otherwise, $\langle\{x, y, z, u, u', u''\}\rangle$ is a replicated-twin C_4, where $u', u'' \in S \setminus N[u]$). In fact, u is adjacent with every vertex in $N(v) \setminus S$.

Now, since $uv' \in E$ and $vv' \notin E$ and since u is the central vertex of the paw or claw $\langle\{u, x, y, z\}\rangle$, the choice of v implies that there must exist a (unique) vertex $w \in S \setminus N[u]$ and that u cannot be adjacent with any vertex in $V \setminus (N[v] \cup \{v'\})$. Hence $N(v') \setminus N(v) = \emptyset$ (otherwise, $\langle\{t, v', u, v\}\rangle$ is a P_4 for $t \in N(v') \setminus N(v)$).

Suppose $w' \in V - (N[v] \cup \{v'\})$. Since G is connected, we may choose w' such that it is adjacent to some vertex w'' in $N(v) \setminus \{u\}$. By the above observations, $w'' \notin S$. So $w'' \in N(v) \setminus S$; but then $\langle\{w', w'', u, v'\}\rangle$ is a P_4. Thus $V \setminus (N[v] \cup \{v'\}) = \emptyset$.

Note that $G - v$ is connected. We can argue as in Subcase 2.1 that there must exist a vertex $u' \in V(G - v)$ which is an extreme vertex in $G - v$ but not in G. Hence u' is the central vertex of a paw or claw $\langle\{u', v, a, b\}\rangle$ containing v. Since $V \setminus (N[v] \cup \{v'\}) = \emptyset$ and $\langle\{v, a, b\}\rangle$ contains at most one edge, we must have (without loss of generality) $a = v'$ and $b \in N(v) \setminus S$, and, hence, $u' \in S$. But then u' is the central vertex of the claw $\langle\{u', x, y, z\}\rangle$ in $G - v$ and hence is not an extreme point of $V(G - v)$, contradicting the choice of u'. ∎

2.2. g_R-convex geometries. In this section we show that if $R = \{k_1, k_2, \ldots, k_t\}$ is a collection of positive integers such that $k_1 < k_2 < \cdots < k_t$ and $k_1 = 2$ or 3, then the class of graphs for which the g_R-convex sets form a convex geometry is precisely the same as the class for which the g_{k_1}-convex sets form a convex geometry. For a connected graph G, let $\mathcal{M}_{g_R}(G)$ be the collection of all g_R-convex sets of G.

THEOREM 10. *Let $G = (V, E)$ be a connected graph and $R = \{k_1, k_2, \ldots, k_t\}$ be a collection of positive integers such that $2 = k_1 < k_2 < \cdots < k_t$. Then the following are equivalent:*

(1) *G is chordal without an induced 3-fan.*

(2) *$(V, \mathcal{M}_{g_R}(G))$ is a convex geometry.*

Proof. Let G be chordal without an induced 3-fan and let S be a g_R-convex set. Then S is g_2- or g-convex since $2 \in R$. Thus, as mentioned in section 1, the extreme vertices of S are the simplicial vertices of $\langle S \rangle$. So, by Theorem 2, S is the convex hull of its extreme points.

For the converse, suppose that G is a connected graph with the property that every g_R-convex set is the convex hull of its extreme points. We show first that G has no induced 3-fan. Suppose G has an induced 3-fan as shown in Figure 2. Let $S = \{u, v, w, x, y\}$. Then S is not g_R-convex, since the extreme points of S are u and x and the g_R-convex hull of $\{u, x\}$ in $\langle S \rangle$ is $\{u, x, y\} \neq S$. Let S' be the g_R-convex hull of S. Then any vertex in $S' \setminus S$ is not an extreme point of S' and thus not simplicial; otherwise, S' is not the smallest convex set containing S. So the extreme points of S' must be contained in $\{u, x\}$. But any proper subset of $\{u, x\}$ does not have a g_R-convex hull that contains S. So u, x are the extreme vertices of S' and thus simplicial in $\langle S' \rangle$. But then the common neighbors of u and x in S' (this includes y) induce a complete graph, and thus these vertices together with u and x form a g_R-convex set which does not contain v or w. So the convex hull of $\{u, x\}$ does not contain S. Thus G has no induced 3-fan. Moreover, G is chordal as we now see. Since V is a g_R-convex set, G must have simplicial vertices. Let v_1 be one of them. Then $V \setminus \{v_1\}$ is g_R-convex and hence either is empty or contains a simplicial vertex. Continuing in this manner, we see that G has a simple elimination ordering and hence is chordal. ∎

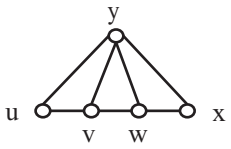


FIG. 2. *The 3-fan.*

THEOREM 11. *Let $G = (V, E)$ be a connected graph and $R = \{k_1, k_2, \ldots, k_r\}$ be a collection of positive integers such that $3 = k_1 < k_2 < \cdots < k_r$. Then the following are equivalent:*

(1) *G is $(P_4, \mathcal{R}_{C_4})$-free.*

(2) *$(V, \mathcal{M}_{g_R}(G))$ is a convex geometry.*

Proof. We show first that if v is a $3SS$ vertex in some g_R-convex set S of G, then v is kSS in $\langle S \rangle$ for all $k \in R$. Suppose $k \in R \setminus \{3\}$ and let $v_1, v_2, \ldots, v_k$ be k distinct vertices of $S \setminus \{v\}$; we need to show that no Steiner tree for $\{v_1, v_2, \ldots, v_k\}$ contains v. Since v is $3SS$, it is not the central vertex of an induced paw, claw, or P_4 in $\langle S \rangle$. Suppose some Steiner tree for $\{v_1, v_2, \ldots, v_k\}$ contains v; among all such trees, let T be one for which $deg_T(v)$ is minimum. Note that $deg_T(v) \geq 2$.

Suppose $deg_T(v) \geq 3$. Since v is not the central vertex of an induced claw, there exists an edge $xy \in E(\langle N_T(v)\rangle)$; but then $T' = (T - xv) \cup xy$ is a Steiner tree for $\{v_1, v_2, \ldots, v_k\}$, which contains v and has $deg_{T'}(v) = deg_T(v) - 1$, contradicting the choice of T. Hence $deg_T(v) = 2$, say, $N_T(v) = \{x, y\}$. Since $k > 2$, we may assume, without loss of generality, that there is a vertex $z \notin \{x, y, v\}$ such that $xz \in E(T)$. As before, by minimality of $deg_T(v)$, $xy, zy \notin E$. Thus, since $\langle\{x, y, z, v\}\rangle$ is not a P_4, $zv \in E$, implying the contradiction that $\langle\{x, y, z, v\}\rangle$ is a paw with v as the central vertex. This shows that the extreme vertices of a g_R-convex set are exactly its $3SS$ vertices.

Now suppose G is $(P_4,\ \mathcal{R}_{C_4})$-free. If S is a g_R-convex set in G, then it is, in particular, a g_3-convex set and hence, by Theorem 9, the g_3-convex hull of its $3SS$ vertices; i.e., S is the g_R-convex hull of its extreme vertices.

Conversely, suppose $(V, \mathcal{M}_{g_R}(G))$ is a convex geometry. Now the same arguments we used in the first paragraph of the proof of Theorem 9 show that G has no induced P_4 or $\mathcal{R}_{C_4}$, since the g_k-convex hull (for each $k \in R$) of any set of at most two vertices is the set itself. □

2.3. g_3^3-convex geometries. Before characterizing the class of graphs for which the g_3^3-convex sets form a convex geometry, we introduce another useful result. Recall that the graphs for which the m^3-convex sets form a convex geometry are characterized in [7] as the (house, hole, domino, A)-free graphs. The proof of this characterization depends on the following result also proven in [7].

THEOREM 12. *If G is a (house, hole, domino, A)-free graph, then every vertex of G either is semisimplicial or lies on an induced path of length at least 3 between two semisimplicial vertices.*

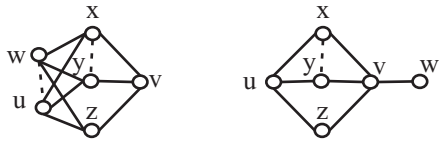


FIG. 3. *Forbidden subgraphs for g_3^3-convex geometries.*

We now proceed to characterize those graphs for which the g_3^3-convex alignment forms a convex geometry. Let $\mathcal{M}_{g_3^3}(G)$ be the collection of all g_3^3-convex sets of a graph G. Recall that a graph F is a *replicated-twin C_4* if it is isomorphic to one of the four graphs shown in Figure 3(a) where any subset of the dotted edges may be chosen to belong to F, and the collection of replicated-twin C_4's is denoted by $\mathcal{R}_{C_4}$. A graph F is a *tailed-twin C_4* if it is isomorphic to one of the two graphs shown in Figure 3(b) where again any subset of the dotted edges may be chosen to belong to F. We denote the collection of tailed-twin C_4's by $\mathcal{T}_{C_4}$.

In order to prove the next main result we first establish the following useful lemma.

LEMMA 13. *Suppose G is a connected distance-hereditary graph that is tailed-twin C_4-free. If G contains an A as an induced subgraph and if u, v are the two leaves of A, then the g_3^3-convex hull of $\{u, v\}$ is precisely $I[u, v]$.*

Proof. Since G is distance-hereditary and the A is an induced subgraph of G, $d(u, v) = 3$. Thus $I[u, v]$ is a subset of the g_3^3-convex hull of $\{u, v\}$. Denote by V_1 the set of vertices at distance 1 from u and by V_2 the set of vertices at distance 2

from u in $I[u,v]$. Necessarily, every vertex in V_2 is adjacent with v. We show first that every vertex in V_1 is adjacent with every vertex in V_2. Suppose some vertex a in V_1 is nonadjacent with some vertex b in V_2. Since $b \in V_2$, there is some $b' \in V_1$ such that $bb' \in E$. Also, since $a \in I[u,v]$, there is some $a' \in V_2$ such that $aa' \in E$. Thus $C = uaa'vbb'u$ is a 6-cycle whose only possible chords are $ab', b'a', a'b$. Thus C is a 6-cycle without crossing chords. By Theorem 4, this is not possible in a distance-hereditary graph. Thus u, v are the only vertices in $I[u,v]$ that are distance (at least) 3 apart. Hence, if there is a vertex in g_3^3-conv$(\{u,v\})$ (denote this set by S) that does not belong to $I[u,v]$, then there is some vertex w in $S \setminus I[u,v]$ such that w is on the Steiner tree of some set of three vertices $\{a,b,c\}$ in $I[u,v]$. Thus $\langle\{a,b,c\}\rangle$ is a disconnected graph. From the above observation, the only sets of three vertices in $I[u,v]$ that induce a disconnected graph are sets that (i) contain both u and v and one vertex from either V_1 or V_2, (ii) contain u and two vertices from V_2, (iii) contain v and two vertices in V_1, or (iv) are contained in V_1 or in V_2. In (i), $I(\{a,b,c\})$ consists of $V_2 \cup \{a,b,c\}$ or $V_1 \cup \{a,b,c\}$, respectively; in (ii), $I(\{a,b,c\})$ consists of $V_1 \cup \{a,b,c\}$; and in (iii), $I(\{a,b,c\})$ consists of $V_2 \cup \{a,b,c\}$. Thus we must be in case (iv), and so $\{a,b,c\}$ is contained in V_1 or V_2. We consider the first case, since the second case can be argued similarly. Let x be any vertex in V_2. By the above observation, a, b, and c are all adjacent with x. Thus $\{u,a,b,c,x,v\}$ induces a tailed-twin C_4 which is forbidden. □

THEOREM 14. *For a connected graph $G = (V,E)$ the following are equivalent:*

(1) *G is (house, hole, domino, A, 3-fan, $\mathcal{R}_{C_4}$, $\mathcal{T}_{C_4}$)-free.*

(2) *$(V, \mathcal{M}_{g_3^3}(G))$ is a convex geometry.*

Proof. To show that (2) implies (1), suppose F is a house, hole, domino, replicated-twin C_4, or tailed-twin C_4. Then F has at most one $3SS$ vertex. Suppose G is a graph that contains F as an induced subgraph. Then the set of extreme points of the convex hull of $V(F)$ is contained in the collection of $3SS$ vertices of F. So the convex hull of the extreme points of the g_3^3-convex hull of $V(F)$ is empty or consists of a single vertex. So in this case the g_3^3-convex alignment of G does not form a convex geometry. Moreover, the 3-fan has two $3SS$ vertices that are distance 2 apart. So the convex hull of the set consisting of these two $3SS$ vertices is just the set itself. Thus if G contains a 3-fan, then the convex hull of the vertices in the 3-fan is not the convex hull of its extreme points. Suppose now that G contains an A. Since G is (house, hole, domino, 3-fan)-free, G is distance-hereditary by Theorem 5. Let the leaves of the A be u and v. Since G is also tailed-twin C_4-free, it follows from Lemma 13 that the g_3^3-convex hull of $\{u,v\}$ consists of the vertices in $I[u,v]$ and hence does not contain the two vertices of A that are not on the u-v geodesic. Let S be the g_3^3-convex hull of the vertices in A. Then its extreme vertices are contained in the set $\{u,v\}$. So the convex hull of the extreme points of S does not contain all the vertices in A and hence is not S. Thus (2) implies (1).

We now show that (1) implies (2). It is not difficult to see that if G is a connected graph of order at most 4, then every g_3^3-convex set is the convex hull of its extreme points. Suppose now that there exists a connected (house, hole, domino, A, 3-fan, $\mathcal{R}_{C_4}$, $\mathcal{T}_{C_4}$)-free graph G (abbreviated by *HHDA* 3-fan $\mathcal{R}_{C_4}\mathcal{T}_{C_4}$-free graph G) for which $(V, \mathcal{M}_{g_3^3})$ is not a convex geometry. We may assume that G is such a graph of smallest possible order. Thus every proper connected induced subgraph of G has the property that its vertex set is the g_3^3-convex hull of its extreme points, i.e., the $3SS$ vertices. By assumption, V is not the g_3^3-convex hull of its extreme points. Since V is g_3^3-convex, it is g^3-convex; so by Corollary 7, V is m^3-convex. By Theorem 12, every vertex of G either is semisimplicial or lies on an induced path of length at least 3 between two

semisimplicial vertices. By Theorem 5, such a path is necessarily a geodesic of length at least 3. Thus if every semisimplicial vertex is $3SS$, then V is the g_3^3-convex hull of its extreme points, which is a contradiction. Let S be the g_3^3-convex hull of $ex(V)$. We now assume that $V \setminus S \neq \emptyset$.

Case 1. $V \setminus S$ contains a vertex a that is not semisimplicial. Since G is *HHDA*-free and V is m^3-convex, Theorem 12 guarantees that a lies on an induced path of length at least 3 between two semisimplicial vertices w, w' of G. By Theorem 5, such an induced w-w' path is a geodesic. Among all pairs $\{w, w'\}$ of semisimplicial vertices such that $a \in I[w, w']$ and $d(w, w') \geq 3$ we will assume that $\{v, v'\}$ is a pair that has a maximum number of $3SS$ vertices. Let $k = d(v, v') \geq 3$. At least one of v and v', say, v, is not $3SS$ in G; otherwise, a lies on a geodesic of length at least 3 between two extreme vertices of V. We may also assume that all neighbors w of v that are also distance k from v' are not $3SS$ vertices; otherwise, a lies on a w-v' geodesic, and we have a contradiction to the choice of the pair $\{v, v'\}$. Since v is semisimplicial but not $3SS$, it must be the central vertex of an induced claw or paw in G. No neighbor of v is distance $k+1$ from v'; otherwise, v is not semisimplicial. So all neighbors of v are distance k or $k-1$ from v'. Let x, y, z be the neighbors of v in a claw or paw and assume z is nonadjacent to both x and y. If one of x, y, or z is distance $k-1$ from v', then all three of these vertices are distance $k-1$ from v'; otherwise, v is not semisimplicial. Suppose x, y, and z are all distance $k-1$ from v'. Then every neighbor of x at distance $k-2$ from v' is a neighbor of z, and every neighbor of y at distance $k-2$ from v' is a neighbor of z; otherwise, v is not semisimplicial. If x and y have a common neighbor at distance $k-2$ from v', then G has a tailed-twin C_4 as an induced subgraph which is forbidden. Since this does not happen, $xy \in E$ and every neighbor of x distance $k-2$ from v' is not adjacent with y. Let w be a neighbor of x distance $k-2$ from v'. Then $\langle\{v, x, y, z, w\}\rangle$ induces a house, which is forbidden.

We may thus assume x, y, and z are all distance k from v'. Note that any neighbor of v that is distance $k-1$ from v' in G is necessarily adjacent with x, y, and z, since v is semisimplicial in G. So a lies on a geodesic of length $k \geq 3$ between each of the vertices in $\{x, y, z\}$ and v'. So x, y, and z are not $3SS$. In particular, either z is the central vertex of an induced claw or paw or z is not semisimplicial.

Subcase 1.1. z is not semisimplicial. Let $P = rzst$ be an induced P_4 containing z as central vertex. Then $d(r, t) = 3$, so $rt \notin E$ and r and t have no common neighbor. Also, t is neither x nor y; otherwise, $rzvt (= x \text{ or } y)$ is an induced P_4 containing v as a central vertex (unless $rv \in E$, which is not possible, since $d(r, t) = 3$). Clearly, $v \neq s$, since v is semisimplicial. Suppose $v = r$. Since $xvzs$ and $yvzs$ are paths of length 3 containing v as central vertex, since z is nonadjacent with both x and y, and since $r (= v)$ is nonadjacent with s, we have $sx, sy \in E$; otherwise, v is not semisimplicial. But then $\langle\{x, y, z, s, t, v (= r)\}\rangle$ is a tailed-twin C_4, which is impossible. So we may assume $\{v, x, y\} \cap \{r, z, s, t\} = \emptyset$.

If $vs \notin E$, then $xs, ys \in E$, since v is semisimplicial. If $rv \in E$, then $rvxs$ and $rvys$ are induced paths of length 3 containing v as the central vertex, unless $rx, ry \in E$. Thus $\langle\{r, v, x, y, z, s\}\rangle$ induces a replicated-twin C_4, which is not possible. So $rv \notin E$. Now $rzvx$ and $rzvy$ are induced P_4's having v as the central vertex, unless $rx, ry \in E$. Again, $\langle\{r, v, x, y, z, s\}\rangle$ induces a replicated-twin C_4, which is forbidden.

So $vs \in E$. If $rv \in E$, then $vt \notin E$, since $d(r, t) = 3$. But then $tsvr$ is an induced P_4 containing v as the central vertex. So $rv \notin E$. Necessarily, $rx, ry \in E$, since v is semisimplicial. Now $rxvs$ and $ryvs$ induce P_4's having v as the central vertex, unless $xs, ys \in E$. But then $\langle\{v, s, x, y, z, r\}\rangle$ induces a replicated-twin C_4, which is forbidden.

Thus we have shown that, whenever v is the central vertex of a claw or paw in an induced subgraph $\langle\{v, x, y, z\}\rangle$ where $zx, zy \notin E$, then z is semisimplicial and thus the central vertex of an induced claw or paw.

Subcase 1.2. z is the central vertex of an induced claw or paw in G. Let r, s, t be the neighbors of z in such a claw or paw. Note that these three vertices differ from x and y, since they are adjacent with z. We may assume that v does not equal r or s.

Subcase 1.2.1. There is an induced claw or paw with z as central vertex that also contains v. Using the above notation we assume $v = t$. We may assume that v is nonadjacent with r. This implies that $xr, yr \in E$, since otherwise $rzva$ is an induced P_4 with v as the central vertex for some $a \in \{x, y\}$. But then $\langle\{r, s, x, y, z, v\}\rangle$ is a replicated-twin C_4, which is forbidden.

If $vs \notin E$, then $xs, ys \in E$; otherwise, $szva$ is an induced P_4 with v as the central vertex for some $a \in \{x, y\}$.

Suppose $vs \in E$. Then $rs \notin E$. Now $svyr$ and $svxr$ are induced P_4's, unless $sx, sy \in E$. Again, $\langle\{r, s, x, y, z, v\}\rangle$ is a replicated-twin C_4, which is forbidden.

Subcase 1.2.2. No induced claw or paw having z as the central vertex contains v. Then $v \notin \{r, s, t\}$ and v and z together with any two of r, s, t do not induce a claw or paw with z as central vertex. In particular, v must be adjacent with at least two of the three vertices r, s, t. Suppose that v is nonadjacent to one of r, s, t, say, r. Then $rs, vs, vt \in E$. So $yvzr$ and $xvzr$ are induced P_4's having v as a central vertex, unless $rx, ry \in E$. Since $rs \in E$ and $\langle\{z, r, s, t\}\rangle$ is a claw or paw with z as the central vertex, $ts, tr \notin E$. Since $vt \in E$, $tvxr$ and $tvyr$ induce P_4's containing v as the central vertex, unless $tx, ty \in E$. But then $\langle\{r, x, y, z, v, t\}\rangle$ is a replicated-twin C_4, which is forbidden.

So v is adjacent with all three vertices r, s, t. Thus $\langle\{v, r, s, t\}\rangle$ induces a claw or paw. We may assume $ts, tr \notin E$. As in Subcase 1.1 we can show that t is semisimplicial and thus the central vertex of an induced claw or paw with vertices t, r_1, s_1, t_1 where we may assume that t_1r_1, t_1s_1 are not edges of G. From Subcase 1.2.1 we may assume $v, z, r, s, t \notin \{r_1, s_1, t_1\}$. From Subcase 1.2.2, both v and z are adjacent with all three vertices in $\{r_1, s_1, t_1\}$. So $x, y \notin \{r_1, s_1, t_1\}$. Now t_1 is semisimplicial and thus the central vertex of some induced claw or paw $\langle\{t_1, r_2, s_2, t_2\}\rangle$ where we may assume $t_2s_2, t_2r_2 \notin E$. Moreover, one can argue as before that $v, z, x, y, r, s, t, r_1, s_1, t_1 \notin \{r_2, s_2, t_2\}$ and that v, z, t, and t_1 are all adjacent with r_2, s_2, t_2. This shows that G has infinitely many vertices, which is not possible. So this case cannot occur.

Case 2. Every vertex of $V \setminus S$ is semisimplicial, and thus each vertex of $V \setminus S$ is the central vertex of a claw or paw.

Subcase 2.1. $diam(G) \leq 2$. By Theorem 5, G is distance hereditary. Thus, since $diam(G) \leq 2$, every vertex of G is semisimplicial, and the extreme points of V are precisely the vertices that are not the central vertex of an induced claw or paw. Moreover, the g_3^3-convex sets of G are precisely the g_3-convex sets of G, and the g_3^3-convex hull of any set of vertices in G is the same as the g_3-convex hull of the set. Since G contains no induced P_4 and is $\mathcal{R}_{C_4}$-free, Theorem 9 implies that V is the g_3^3-convex hull of its extreme points. So this case cannot occur.

Subcase 2.2. $diam(G) \geq 3$. Then G has vertices that are not semisimplicial. These necessarily belong to S. Thus S must contain at least two vertices that are not semisimplicial. Each of these must either be central vertices of a geodesic between two vertices of S that are distance at least 3 apart or belong to a Steiner tree of a set of three vertices of S. Thus S has at least four vertices and is thus connected (since it contains the Steiner interval of every subset of three vertices that it contains).

Observe that $G - S$ has only one component; for if $H_1, H_2, \ldots, H_k$ are the compo-

nents of $G - S$ where $k \geq 2$, then $G - V(H_k)$ is connected, and thus, by our choice of G, $V \setminus V(H_k)$ is the g_3^3-convex hull of its extreme points. No vertex of H_k is adjacent with any vertices of H_i for $1 \leq i \leq k-1$; thus the vertices of H_i for $1 \leq i \leq k-1$ are still the central vertex of a claw or paw in $G - V(H_k)$. Therefore, the extreme points of $V \setminus V(H_k)$ are contained in S. However, the g_3^3-convex hull of any subset of vertices of S is contained in S. Hence the g_3^3-convex hull of the extreme points of $V \setminus V(H_k)$ is not $V \setminus V(H_k)$, contrary to our choice of G. Thus $G - S$ has exactly one component, say, H.

Observe that $diam(H) \leq 2$, since G is distance-hereditary and every vertex of H is semisimplicial.

Subcase 2.2.1. H contains a vertex v whose eccentricity in G is at least 3. Let v' be an eccentric vertex for v. Then $d = d(v, v') \geq 3$. Since $diam(H) \leq 2$, we have $v' \in S$. Let V_i be the collection of vertices distance i from v in $I[v, v']$ $(1 \leq i \leq d)$. Since the vertices of $V_1 \cup V_2$ are not semisimplicial, they belong to S.

We show that neighbors of H that belong to S are either in V_1 or at distance at least d from v'. Let w be a neighbor of v in $S \setminus V_1$. If $d(w, v') \leq d-1$, then either $w \in V_1$ or $d(v, v') < d$, neither of which is possible. So $d(w, v') \geq d$. Suppose now that u is a neighbor of v in H. Then u is not adjacent with a vertex of V_i for $i \geq 2$; otherwise either u is not semisimplicial or $d(v, v') < d$, neither of which is possible. Since v is semisimplicial, u is adjacent with every vertex of V_1. So $d(u, v') \leq d$. If $d(u, v') < d$, then either u is not semisimplicial or $d(v, v') < d$, which is not possible. Hence $d(u, v') = d$ for every neighbor u of v in H. As we argued for v, the neighbors of u in S either are in V_1 or are distance at least d from v'. Since $diam(H) \leq 2$, every vertex of H not adjacent with v (if any) is necessarily adjacent with a neighbor of v. Suppose x is a vertex of H distance 2 from v. Let u be a common neighbor of v and x in H. Since $d(u, v') = d$, we can argue as above that every neighbor of x in S is either in V_1 or at distance at least d from v'.

Now v' is not a cut-vertex of G, since it is an eccentric vertex for v. So, by our choice of G, the g_3^3-convex hull of the extreme points of $V(G - v')$ is $V(G - v')$. Since no vertex of H is adjacent with v', each vertex of H is the central vertex of a claw or paw in $G - v'$. So the extreme points of $V(G - v')$ are contained in S. From Theorem 8 it follows that $S \setminus \{v'\}$ is a g_3^3-convex set, and the convex hull of the extreme points of $V(G - v')$ is thus contained in $S \setminus \{v'\}$, which is a contradiction.

Subcase 2.2.2. Every vertex of H has eccentricity 2 in G. (Note that since $diam(G) \geq 3$, no vertex of G has eccentricity 1.) Then S contains all the vertices whose eccentricity equals the diameter. If v is a vertex of H, then v is not adjacent with a vertex having eccentricity at least 3; otherwise, v either is not semisimplicial or has eccentricity at least 3, neither of which is possible. Thus no vertex of G having eccentricity equal to the diameter is adjacent with a vertex of H. Let w be a vertex such that $ecc(w) = diam(G)$. Then w is not a cut-vertex of G. So by the choice of G, the g_3^3-convex hull of the extreme points of $V(G - w)$ is $V(G - w)$. Since no vertex of H is adjacent with w, the extreme points of $V(G - w)$ are contained in $S \setminus \{w\}$. Since $S \setminus \{w\}$ is a g_3^3-convex set of $G - w$, the g_3^3-convex hull of the extreme points of $V(G - w)$ is thus contained in $S \setminus \{w\}$, which is a contradiction. □

3. Concluding remarks. It appears to be an open problem to determine the class of graphs for which the g^3-convex sets form a convex geometry. Moreover, the class of graphs for which every LexBFS ordering is a weakly semisimplicial ordering has not yet been characterized, and neither has the class of graphs for which every MCS ordering of every induced subgraph is weakly semisimplicial. The task of char-

acterizing extreme vertices of g_k-convex sets for $k \geq 4$ will become increasingly more tedious and thus less appealing. However, in view of section 2.2, the problem of examining relationships between classes of graphs for which certain Steiner convexities are convex geometries is more interesting. Suppose $R = \{k_1, k_2, \ldots, k_t\}$ is a collection of positive integers such that $2 \leq k_1 < k_2 < \cdots < k_t$. One may, for example, ask whether the class of graphs for which the g_R-convex sets form a convex geometry is the same as the class of graphs for which the g_{k_1}-convex sets form a convex geometry. Of course, this question was answered in the affirmative in section 2.2 for the special cases where $k_1 = 2$ and $k_1 = 3$.

REFERENCES

[1] H.-J. Bandelt and H. M. Mulder, *Distance-hereditary graphs*, J. Combin. Theory Ser. B, 41 (1986), pp. 182–208.

[2] J. M. Bilbao and P. H. Edelman, *The Shapley value on convex geometries*, Discrete Appl. Math., 103 (2000), pp. 33–40.

[3] A. Brandstädt, V. B. Le, and J. P. Spinrad, *Graph Classes: A Survey*, SIAM Monographs on Discrete Mathematics and Applications 3, SIAM, Philadelphia, 1999.

[4] J. Caceres and O. R. Oellermann, *On 3-Steiner Simplicial Elimination*, Discrete Math., to appear.

[5] G. Chartrand and L. Lesniak, *Graphs and Digraphs*, 3rd ed., Chapman and Hall, New York, 1996.

[6] D. P. Day, O. R. Oellermann, and H. C. Swart, *Steiner distance-hereditary graphs*, SIAM J. Discrete Math., 7 (1994), pp. 437–442.

[7] F. F. Dragan, F. Nicolai, and A. Brandstädt, *Convexity and HHD-free graphs*, SIAM J. Discrete Math., 12 (1999), pp. 119–135.

[8] M. Farber and R. E. Jamison, *Convexity in graphs and hypergraphs*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 433–444.

[9] E. Howorka, *A characterization of distance hereditary graphs*, Quart. J. Math. Oxford Ser. (2), 28 (1977), pp. 417–420.

[10] B. Jamison and S. Olariu, *On the semi-perfect elimination*, Adv. in Appl. Math., 9 (1988), pp. 364–376.

[11] E. Kubicka, G. Kubicki, and O. R. Oellermann, *Steiner intervals in graphs*, Discrete Appl. Math., 81 (1998), pp. 181–190.

[12] O. R. Oellermann, *Convexity Notions in Graphs*, http://www-ma2.upc.edu/seara/wmcgt06/.

[13] O. R. Oellermann and M. L. Puertas, *Steiner intervals and Steiner geodetic numbers in distance hereditary graphs*, Discrete Math., 307 (2007), pp. 88–96.

[14] D. J. Rose, R. E. Tarjan, and G. S. Leuker, *Algorithmic aspects of vertex elimination on graphs*, SIAM J. Comput., 5 (1976), pp. 266–283.

[15] 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.

[16] M. J. L. Van de Vel, *Theory of Convex Structures*, North–Holland, Amsterdam, 1993.