SIAM J. Discrete Math.
Vol. 17, No. 1, pp. 1–17

SOME NEW ASPECTS OF THE COUPON COLLECTOR'S PROBLEM[*]

AMY N. MYERS[†] AND HERBERT S. WILF[‡]

Abstract. We extend the classical coupon collector's problem to one in which two collectors are simultaneously and independently seeking collections of d coupons. We find, in finite terms, the probability that the two collectors finish at the same trial, and we find, using the methods of Gessel–Viennot, the probability that the game has the following "ballot-like" character: the two collectors are tied with each other for some initial number of steps, and after that the player who first gains the lead remains ahead throughout the game. As a by-product we obtain the evaluation in finite terms of certain infinite series whose coefficients are powers and products of Stirling numbers of the second kind.

We study the variant of the original coupon collector's problem in which a single collector wants to obtain at least h copies of each coupon. Here we give a simpler derivation of results of Newman and Shepp and extend those results. Finally we obtain the distribution of the number of coupons that have been obtained exactly once ("singletons") at the conclusion of a successful coupon collecting sequence.

Key words. coupon, Stirling number, lattice path

AMS subject classification. 05-02

DOI. 10.1137/S0895480102403076

1. Introduction and results.

The classical coupon collector's problem is the following. Suppose that a breakfast cereal manufacturer offers a souvenir ("coupon") hidden in each package of cereal, and there are d different kinds of souvenirs altogether. The collector wants to have a complete collection of all d souvenirs. What is the probability $p(n, d)$ that exactly n boxes of cereal will have to be purchased in order to obtain, for the first time, a complete collection of at least one of each of the d kinds of souvenir coupons?

The answer to that question is well known (e.g., [5, p. 132]) to be

$$p(n,d) = \frac{d!}{d^n}\left\{{n-1 \atop d-1}\right\}, \tag{1.1}$$

where the $\left\{{n \atop k}\right\}$'s are the Stirling numbers of the second kind.

We study, in this paper, a number of other aspects of this problem as well as a generalization of it to a two-player game.

First, suppose we have two coupon collectors, drawing coupons simultaneously, and each seeking to obtain a complete collection of d coupons. We ask for the probability that the two games are completed at the same time. The answer is given by (2.6) below. That answer is expressed in finite terms, owing to the closed form evaluation of the ordinary power series generating function for the squares of the Stirling numbers of the second kind, contained in (2.5).

[*] Received by the editors February 25, 2002; accepted for publication (in revised form) December 20, 2002; published electronically September 17, 2003.
http://www.siam.org/journals/sidma/17-1/40307.html

[†] Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104-6395. Current address: Mathematics and Computer Science Department, St. Joseph's University, Philadelphia, PA 19131 (amyers@sju.edu).

[‡] Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104-6395 (wilf@central.cis.upenn.edu).

Next, we consider the following two-person game. Again two coupon collectors are simultaneously drawing coupons at random. This time we are interested in a ballot-like problem: What is the probability that the player who first completed a collection (the *winner*) was never behind (i.e., never had fewer distinct coupons) at any intermediate stage of the play? Here we give a complete answer to a slightly easier question, namely the following: What is the probability that, after an initial segment of play in which the players are tied, one of them takes the lead and keeps the lead strictly until the end. The answer is in (2.26) below and is obtained by the Gessel–Viennot theory of nonintersecting lattice paths.

In each of these cases the answer can first be written as an infinite series whose coefficients involve various products of Stirling numbers. What is interesting, though, is that in all such cases we are able to express the answers in finite terms. Indeed, one of our main results here is the observation that infinite series whose coefficients involve various powers and products of Stirling numbers of the second kind can readily be evaluated in finite terms.

In section 4 we return to the original collecting problem of obtaining at least one copy of each coupon, but now we study the variant of the problem in which a single collector wants to obtain at least $h \geq 1$ copies of each coupon. We obtain the generating function (4.5) for the probability that exactly n trials are needed, the exact value of the average number of trials (4.10), and the asymptotic behavior (4.15) of these quantities as $n \to \infty$.

Finally, in section 5 we study the number of coupons that have been collected only once, at the end of a collection sequence. We find the distribution function (5.4) for this number and show that the average number of these singletons is just the harmonic number $H_d = 1 + 1/2 + \cdots + 1/d$.

2. The two-person collecting competition.

2.1. Simultaneous completion. We find now the probability of simultaneous completion of two independent coupon collecting sequences. Evidently this is

$$\sum_{n\geq 0} p(n,d)^2 = \sum_{n\geq 0} \frac{d!^2}{d^{2n}} \left\{ {n-1 \atop d-1} \right\}^2, \tag{2.1}$$

which expresses the answer as an infinite sum. We can rewrite this as a finite sum by finding a finite expression for the generating function for the squares of the Stirling numbers of the second kind,

$$F_k(x) \overset{\text{def}}{=} \sum_{n\geq k} \left\{ {n \atop k} \right\}^2 x^n,$$

analogously to the well-known generating function for these numbers themselves,

$$\sum_{n\geq k} \left\{ {n \atop k} \right\} x^n = \frac{x^k}{(1-x)(1-2x)\dots(1-kx)}. \tag{2.2}$$

The easiest way to do this is via the standard explicit formula for these Stirling numbers, viz.

$$\begin{aligned} \left\{ {n \atop k} \right\} &= \frac{1}{k!}\sum_{r=1}^{k}(-1)^{k-r}\binom{k}{r} r^n \qquad (1 \leq k \leq n) \\ &\overset{\text{def}}{=} \sum_{r=1}^{k} A_{k,r} r^{n-k}, \end{aligned} \tag{2.3}$$

where we have written

$$A_{k,r} = \frac{(-1)^{k-r} r^k}{k!} \binom{k}{r}. \tag{2.4}$$

It follows that

$$\begin{aligned} F_k(x) \overset{\text{def}}{=} \sum_{n \geq k} {n \brace k}^2 x^n &= \sum_{n \geq k} x^n \sum_{r,s=1}^{k} A_{k,r} A_{k,s} r^{n-k} s^{n-k} \\ &= x^k \sum_{r,s=1}^{k} A_{k,r} A_{k,s} \sum_{n \geq k} (rsx)^{n-k} \\ &= x^k \sum_{r,s=1}^{k} \frac{A_{k,r} A_{k,s}}{1 - rsx} \qquad \left(|x| < \frac{1}{k^2} \right). \end{aligned} \tag{2.5}$$

Thus for the simultaneous completion probability we obtain, from (2.1),

$$\sum_{n \geq 0} p(n,d)^2 = \frac{d!^2}{d^{2d}} \sum_{r,s=1}^{d-1} \frac{A_{d-1,r} A_{d-1,s}}{1 - \frac{rs}{d^2}} \tag{2.6}$$

by (2.5), where the A's are given by (2.4). This sequence of probabilities, for $d = 1, 2, \ldots$, begins as

$$1, \frac{1}{3}, \frac{11}{70}, \frac{9}{91}, \frac{688877}{9561123}, \frac{358555}{6330324}, \frac{2730269557627901}{58560931675094420}, \frac{146271649897951}{3695016639410525}, \ldots,$$

i.e., as

$$1, 0.33333\ldots, 0.15714\ldots, 0.098901\ldots, 0.072049\ldots,$$
$$0.056640\ldots, 0.046622\ldots, 0.039586\ldots, \ldots.$$

2.2. Neck-and-neck then always ahead. We encode a sequence of n draws as a path ω with n vertices in the lattice $\mathcal{L}$ consisting of vertices (i,j) and edges $\{(i,j),(i+1,j)\}, \{(i,j),(i+1,j+1)\}$ for all $i,j \geq 0$. The first coordinate of a vertex in the path gives the number of draws, or *steps*, and the second coordinate gives the number of distinct coupons the collector has at that step. Thus ω starts at $(0,0)$ indicating the collector has 0 coupons at draw 0, proceeds to $(1,1)$ (the collector has 1 coupon after 1 draw), and ends at (n,d), $n \geq d$ (the collector has a complete collection at step n). We write $\omega = (0,0)\overline{\omega}(n,d)$, where $\overline{\omega}$ is a path from $(1,1)$ to $(n-1, d-1)$, to indicate that ω starts at the vertex $(0,0)$, continues with the first vertex $(1,1)$ in $\overline{\omega}$, then follows $\overline{\omega}$ through to $(n-1,d-1)$, and finally ends with the vertex (n,d).

We assign a weight of i/d to each horizontal edge $\{(i,j),(i+1,j)\}$ in the lattice $\mathcal{L}$. This is the probability that at the $(j+1)$st step, the collector draws one of the i distinct coupons already collected at step j. We assign a weight of $1 - i/d$ to each *northeast* edge $\{(i,j),(i+1,j+1)\}$. The probability that the collector draws the particular sequence of coupons encoded by the path ω is given by the product of the weights on the edges of ω. We let $P(\omega)$ denote this probability.

Suppose one collector, the *winner*, collects all d distinct coupons for the first time at step n. (At step $n-1$ the winner had $d-1$ distinct coupons.) Let ω_1 be the lattice path which encodes the winner's sequence of draws. Let ω_2 encode the other collector's draws. We compute the probability $p(d)$ that ω_1 and ω_2 are identical until some point at which the winner takes the lead and the other collector never catches up.

To do this, we begin by supposing ω_1 is identical to ω_2 until step k, at which point both collectors have d_1 distinct coupons. The argument splits into two cases, namely $k \le n-2$ and $k = n-1$. In both cases, at step $k+1$ the winner collects one additional distinct coupon while the other collector does not. After step k, the two paths never intersect again. The winner collects all d distinct coupons for the first time at step n. Suppose the other collector has d_2 distinct coupons at this point. The probability we seek is

$$p(d) = \sum_{n=d}^{\infty} \sum_{k=1}^{n-1} \sum_{d_1=1}^{d-1} \sum_{d_2=d_1}^{d-1} \sum_{(\omega_1,\omega_2)} P(\omega_1)P(\omega_2), \tag{2.7}$$

where the innermost sum ranges over all pairs (ω_1, ω_2) described above.

2.3. The case $k \le n-2$. Write $\omega_1 = \alpha\overline{\omega_1}(n,d)$, where α denotes a lattice path from $(0,0)$ to (k, d_1), and $\overline{\omega_1}$ denotes a path from $(k+1, d_1+1)$ to $(n-1, d-1)$. Similarly, set $\omega_2 = \alpha\overline{\omega_2}$, where α is as above and $\overline{\omega_2}$ is a path from $(k+1, d_1)$ to (n, d_2). Note that $\overline{\omega_1}$ and $\overline{\omega_2}$ are nonintersecting paths in the lattice $\mathcal{L}$. In terms of these we have $P(\omega_1) = P(\alpha)(1 - d_1/d)P(\overline{\omega_1})(1/d)$ and $P(\omega_2) = P(\alpha)(d_1/d)P(\overline{\omega_2})$. Hence from (2.7) we find for the combined probability of all pairs if $k \le n-2$,

$$\begin{aligned} p(d)_{k\le n-2} &= \sum_{n=d}^{\infty} \sum_{k=1}^{n-2} \sum_{d_1=1}^{d-1} \sum_{d_2=d_1}^{d-1} \sum_{(\omega_1,\omega_2)} \left[P(\alpha)\left(1-\frac{d_1}{d}\right)P(\overline{\omega_1})\left(\frac{1}{d}\right)\right]\left[P(\alpha)\left(\frac{d_1}{d}\right)P(\overline{\omega_2})\right] \\ &= \sum_{n=d}^{\infty} \sum_{k=1}^{n-2} \sum_{d_1=1}^{d-1} \sum_{d_2=d_1}^{d-1} \left(1-\frac{d_1}{d}\right)\left(\frac{1}{d}\right)\left(\frac{d_1}{d}\right)\sum_{\alpha} P(\alpha)^2 \sum_{(\overline{\omega_1},\overline{\omega_2})} P(\overline{\omega_1})P(\overline{\omega_2}). \end{aligned} \tag{2.8}$$

At this point we have translated a question about coupon collecting into a problem involving nonintersecting paths in a lattice. We have set the stage for application of the Gessel–Viennot theorem [3]. This result concerns pairs of nonintersecting lattice paths with no constraints on vertices or edges in the paths. For this reason we have written ω_1 and ω_2 in terms of $\overline{\omega_1}$ and $\overline{\omega_2}$.

The theorem refers to an arbitrary set $\mathcal{L}$, which we will take to be the lattice defined earlier, and a weight (or valuation) v, which we take to be P. The theorem equates a sum of weights of paths with the determinant of a matrix $(a_{ij})_{1\le i,j\le l}$. The entries of this matrix are defined by $a_{ij} = \sum_{\omega} v(\omega)$, where ω ranges over all paths from A_i to B_j.

The theorem requires that two given sequences, $(A_1, A_2, \ldots, A_l)$ and $(B_1, B_2, \ldots, B_l)$, of vertices in $\mathcal{L}$, the sets Ω_{ij}, $1 \le i, j \le l$, of all paths in $\mathcal{L}$ between A_i and B_j, and the weight v satisfy both the *finiteness* and *crossing conditions*. The *finiteness condition* requires the set of paths in Ω_{ij} with nonzero weight be finite. The *crossing condition* requires that paths in $\Omega_{ij'}$ and $\Omega_{i'j}$, $i < i'$ and $j < j'$, with nonzero weight share a common vertex. Both conditions hold for the paths we consider.

THEOREM 2.1 (Gessel–Viennot). *Suppose $\mathcal{L}$, v, $(A_1, A_2, \ldots, A_l)$, and $(B_1, B_2, \ldots, B_l)$ satisfy both the finiteness and crossing conditions. Then the determinant of the matrix $(a_{ij})_{1\le i,j\le l}$ is the sum of the weights of all configurations of paths $(\omega_1, \omega_2, \ldots, \omega_l)$ satisfying the following two conditions:*

(i) *The paths ω_k are pairwise nonintersecting, and*

(ii) *ω_k is a path from A_k to B_k.*

In other words,

$$\det\left(\{a_{ij}\}_{i,j=1}^{l}\right) = \sum_{(\omega_1,\omega_2,\dots,\omega_l)} v(\omega_1)v(\omega_2)\dots v(\omega_l).$$

Application of this theorem to our problem requires the computation of only a 2×2 determinant! Let $A_1 = (k+1, d_1+1)$, $A_2 = (k+1, d_1)$, $B_1 = (n-1, d-1)$, and $B_2 = (n, d_2)$. Then

$$\sum_{(\overline{\omega_1},\overline{\omega_2})} P(\overline{\omega_1})P(\overline{\omega_2}) = \det\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \tag{2.9}$$

where a_{ij} is the sum $\sum_\omega P(\omega)$ over all paths ω from A_i to B_j.

2.4. Paths from *A* to *B*. In this section we compute the probability $P(\omega)$ of an arbitrary path ω from a vertex $A = (a_1, b_1)$ to a vertex $B = (a_2, b_2)$ as well as the sum over all such paths. Such a path contains $b_2 - b_1$ *northeast* edges $\{(i,j),(i+1,j+1)\}$ and $(a_2 - a_1) - (b_2 - b_1)$ horizontal edges. The weights assigned to northeast edges in order from left to right are $1 - \frac{b_1}{d}, 1 - \frac{b_1+1}{d}, \dots, 1 - \frac{b_2-1}{d}$. The weight assigned to a horizontal edge depends on its coordinates. Consider the edge $\{(i,j),(i+1,j)\}$. This edge indicates the collector has j distinct coupons at step i and draws one of the same j coupons at step $i+1$. The probability of this (weight of the edge) is $\frac{j}{d}$. Thus the probability of a path ω from A to B is

$$\begin{aligned} P(\omega) &= \left(\frac{b_1}{d}\right)^{e_1}\left(1 - \frac{b_1}{d}\right)\left(\frac{b_1+1}{d}\right)^{e_2}\left(1 - \frac{b_1+1}{d}\right)\dots\left(1 - \frac{b_2-1}{d}\right)\left(\frac{b_2}{d}\right)^{e_{b_2-b_1+1}} \\ &= \frac{1}{d^{a_2-a_1}}\frac{(d-b_1)!}{(d-b_2)!}(b_1)^{e_1}(b_1+1)^{e_2}\dots(b_2)^{e_{b_2-b_1+1}}, \end{aligned} \tag{2.10}$$

where $e = (e_1, e_2, \dots, e_{b_2-b_1+1})$ is an ordered partition, a *composition*, of $(a_2 - a_1) - (b_2 - b_1)$ into $b_2 - b_1 + 1$ nonnegative integer parts. With this we compute the sum of the probabilities of all paths from A to B.

$$\begin{aligned} \sum_{\omega = A\cdots B} P(\omega) &= \sum_e \left(\frac{b_1}{d}\right)^{e_1}\left(1 - \frac{b_1}{d}\right)\left(\frac{b_1+1}{d}\right)^{e_2}\left(1 - \frac{b_1+1}{d}\right) \\ &\qquad \dots\left(1 - \frac{b_2-1}{d}\right)\left(\frac{b_2}{d}\right)^{e_{b_2-b_1+1}} \\ &= \frac{1}{d^{a_2-a_1}}\frac{(d-b_1)!}{(d-b_2)!}\sum_e (b_1)^{e_1}(b_1+1)^{e_2}\dots(b_2)^{e_{b_2-b_1+1}}, \end{aligned} \tag{2.11}$$

where the sum is over all compositions $e = (e_1, e_2, \dots, d_{b_2-b_1+1})$ of $(a_2 - a_1) - (b_2 - b_1)$ into $b_2 - b_1 + 1$ nonnegative integer parts. This is the coefficient of $x^{(a_2-a_1)-(b_2-b_1)}$ in the series expansion of

$$\frac{1}{(1 - b_1 x)(1 - (b_1+1)x)\dots(1 - b_2 x)},$$

so we can find a simpler formula for it by looking at the partial fraction expansion

$$\frac{1}{\prod_{m=a}^{b}(1-mx)} = \sum_{m=a}^{b} \frac{B_m}{1-mx}, \tag{2.12}$$

where

$$B_m = \frac{(-1)^{b-m} m^{b-a}}{(b-a)!} \binom{b-a}{m-a}.$$

From this and (2.11) we obtain

$$\begin{aligned}\sum_{\omega=A\cdots B} P(\omega) &= \frac{1}{d^{a_2-a_1}} \frac{(d-b_1)!}{(d-b_2)!} [x^{(a_2-a_1)-(b_2-b_1)}] \left\{ \sum_{m=b_1}^{b_2} \frac{B_m}{1-mx} \right\} \\ &= \frac{1}{d^{a_2-a_1}} \frac{(d-b_1)!}{(d-b_2)!} \sum_{m=b_1}^{b_2} B_m m^{(a_2-a_1)-(b_2-b_1)} \\ &= \frac{1}{d^{a_2-a_1}} \frac{(d-b_1)!}{(d-b_2)!} \sum_{m=b_1}^{b_2} \frac{(-1)^{b_2-m} m^{a_2-a_1}}{(b_2-b_1)!} \binom{b_2-b_1}{m-b_1}. \qquad (2.13)\end{aligned}$$

2.5. Evaluating the determinant. We use the results of the previous section to evaluate the determinant in (2.9). To compute a_{11}, we substitute $A = A_1 = (k+1, d_1+1)$ and $B = B_1 = (n-1, d-1)$ into (2.13). This yields

$$a_{11} = \frac{1}{d^{n-k-2}} (d-d_1-1)! \sum_{m=d_1+1}^{d-1} \frac{(-1)^{d-m-1} m^{n-k-2}}{(d-d_1-2)!} \binom{d-d_1-2}{m-d_1-1}. \tag{2.14}$$

In a similar manner we obtain

$$a_{12} = \frac{1}{d^{n-k-1}} \frac{(d-d_1-1)!}{(d-d_2)!} \sum_{m=d_1+1}^{d_2} \frac{(-1)^{d_2-m} m^{n-k-1}}{(d_2-d_1-1)!} \binom{d_2-d_1-1}{m-d_1-1}, \tag{2.15}$$

$$a_{21} = \frac{1}{d^{n-k-2}} (d-d_1)! \sum_{m=d_1}^{d-1} \frac{(-1)^{d-m-1} m^{n-k-2}}{(d-d_1-1)!} \binom{d-d_1-1}{m-d_1}, \tag{2.16}$$

$$a_{22} = \frac{1}{d^{n-k-1}} \frac{(d-d_1)!}{(d-d_2)!} \sum_{m=d_1}^{d_2} \frac{(-1)^{d_2-m} m^{n-k-1}}{(d_2-d_1)!} \binom{d_2-d_1}{m-d_1}. \tag{2.17}$$

Using (2.14)–(2.17) we compute the determinant of our 2 × 2 matrix.

$$\begin{aligned}\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} &= \frac{(d-d_1)!(d-d_1-1)!}{d^{2n-2k-3}(d-d_2)!} \qquad (2.18) \\ &\times \sum_{l=d_1}^{d} \sum_{m=d_1}^{d_2} \frac{m(lm)^{n-k-2}(-1)^{d_1+d_2}(l-d)(m^2-l^2)}{(d-d_1-1)!(d_2-d_1)!} \binom{d-d_1-1}{l-d_1} \binom{d_2-d_1}{m-d_1} \\ &\overset{\text{def}}{=} \det(d, d_1, d_2, k, n). \qquad (2.19)\end{aligned}$$

Substituting (2.19) in (2.9), we obtain

$$\sum_{(\overline{\omega_1}, \overline{\omega_2})} P(\overline{\omega_1}) P(\overline{\omega_2}) = \det(d, d_1, d_2, k, n). \tag{2.20}$$

2.6. The initial common segment. In the previous section we evaluated the determinant in (2.9). In this section we compute the sum $\sum_\alpha P(\alpha)^2$ in (2.8). Recall α is a path from $(0,0)$ to (k,d_1).

Equation (2.10) gives the probability of an arbitrary path from A to B. Substituting $A=(0,0)$ and $B=(k,d_1)$ gives the probability

$$P(\alpha)=\frac{d!}{d^k(d-d_1)!}1^{e_1}2^{e_2}\cdots {d_1}^{e_{d_1}}$$

of an arbitrary path α from $(0,0)$ to (k,d_1). It follows that

$$\begin{aligned}\sum_{\alpha=(0,0)\cdots(k,d_1)} P(\alpha)^2 &= \frac{d!^2}{d^{2k}(d-d_1)!^2}\sum_{e_1+\cdots+e_{d_1}=k-d_1} 1^{2e_1}2^{2e_2}\ldots d_1^{2e_{d_1}}\\ &= \frac{d!^2}{d^{2k}(d-d_1)!^2}[x^{k-d_1}]\left\{\frac{1}{(1-1^2x)(1-2^2x)\ldots(1-d_1^2x)}\right\}\\ &= \frac{d!^2}{d^{2k}(d-d_1)!^2}\sum_{m=1}^{d_1} C_m m^{2k-2d_1} \qquad \text{(as in (2.12))}\\ &= \frac{2d!^2}{(d-d_1)!^2d^{2k}(2d_1)!}\sum_{m\geq 1}(-1)^{d_1-m}\binom{2d_1}{d_1+m}m^{2k} \qquad (2.21)\\ &\overset{\text{def}}{=} \operatorname{init}(d,d_1,k). \qquad (2.22)\end{aligned}$$

2.7. The case $k=n-1$. Suppose now that the two walks are identical up to the point $(n-1,d-1)$. Since step n is the finish, the next step for the winning player will be to (n,d) and for the losing player to $(n,d-1)$. These last steps have respective probabilities $1/d$ and $1-1/d$. Hence the probability of the complete pair of walks in this case is the probability of two identical walks from $(0,0)$ to $(n-1,d-1)$ (which is given by (2.21) with $(k,d_1):=(n-1,d-1)$) multiplied by $(d-1)/d^2$.

2.8. Putting it together. We now substitute (2.19) and (2.22) into (2.8) to obtain the probability of all pairs of paths that we are considering,

$$\begin{aligned}p(d) &= \sum_{n=d}^{\infty}\sum_{k=1}^{n-2}\sum_{d_1=1}^{d-1}\sum_{d_2=d_1}^{d-1}\left(1-\frac{d_1}{d}\right)\left(\frac{d_1}{d}\right)\left(\frac{1}{d}\right)\operatorname{init}(d,d_1,k)\det(d,d_1,d_2,k,n)\\ &\quad + \frac{d-1}{d^2}\sum_{n=d}^{\infty}\operatorname{init}(d,d-1,n-1) \qquad (2.23)\\ &\overset{\text{def}}{=} \Sigma_1+\Sigma_2. \qquad (2.24)\end{aligned}$$

It turns out that the sums over the indices d_2,n,k can all be carried out in explicit closed form. Hence we can obtain an expression which is in finite terms for the total probability.

First, the sum on d_2 in Σ_1 above can be done in closed form since

$$\sum_{d_2=d_1}^{d-1}(-1)^{d_2}\binom{d-d_1}{d-d_2}\binom{d_2-d_1}{t-d_1}=(-1)^{d+1}\binom{d-d_1}{d-t}.$$

Next, the remaining sum over the indices n and k in the first summation Σ_1 is

$$\psi(d,r,s,t) \overset{\text{def}}{=} \sum_{n=d}^{\infty}\sum_{k=1}^{n-2} \frac{r^{2k}t^{n-k-1}s^{n-k-2}}{d^{2n}} = \begin{cases} \frac{r^{2d}(d^3-2d^2-r^2d+3r^2)}{d^{2d-2}(d^2-r^2)^2s^2t} & \text{if } r^2=st, \\ \frac{r^4(st)^{d-1}(r^2-d^2)+str^{2d}(d^2-st)}{d^{2d-2}(d^2-st)(d^2-r^2)(r^2-st)r^2s} & \text{otherwise.} \end{cases} \tag{2.25}$$

The sum over n in Σ_2 is trivial, and so there remain no infinite sums in our final expression for the probability $p(d)$, which is

$$\begin{aligned} &\sum_{d_1=1}^{d-1} \frac{2d!^2 d_1(d-d_1)}{(d-d_1)!^2(2d_1)!} \sum_{r,s,t\geq 1} (-1)^{d_1-r-s-t}(s-t)\binom{2d_1}{d_1+r} \\ &\qquad\qquad \binom{d-d_1-1}{s-d_1}\binom{d-d_1}{d-t}\psi(d,r,s,t) \\ &\qquad + \frac{4(d-1)d!^2}{d^{2d-2}(2d-2)!}\sum_{r=1}^{d-1}(-1)^{d-1-r}\binom{2d-2}{d-1+r}\frac{r^{2d-2}}{d^2-r^2} + \delta_{d,1}, \end{aligned} \tag{2.26}$$

where ψ is given by (2.25).

This is the probability that the game is of the type we described, namely where the players are tied for some initial segment of trials and then the player who pulls ahead remains ahead always, expressed as a finite sum (albeit a complicated one!). More precisely, the values of $p(d)$ can be calculated, as rational numbers, with $O(d^4)$ evaluations of the above summand. The exact values of $p(d)$ for $d = 1, 2, 3, 4, 5, \ldots$ are

$$\left\{1, \frac{2}{3}, \frac{43}{70}, \frac{986}{2275}, \frac{5672893}{1912246}, \ldots\right\}.$$

As decimals, the values of $\{p(d)\}_{d=1}^{10}$ are

$$\{1.0, 0.66667, 0.61429, 0.43341, 0.29667, 0.21177, 0.16016, 0.12748, 0.10551, 0.08988\}.$$

2.9. One collector never behind. In contrast to the problem of staying ahead as soon as the tie is broken, which we have solved in the preceding sections, the problem in which the ultimate winner has never been behind is unsolved.

Suppose the winner collects all d distinct coupons for the first time at step n, at which point the other collector has $d' < d$ distinct coupons. We discuss the probability $b(d)$ the winner has never been behind. We use b for "ballot" since this version of the problem has a distinct *ballot-problem* flavor (see, e.g., [1]).

Let w_1 be the lattice path which encodes the winner's sequence of draws. Let ω_2 encode the other collector's sequence of draws. Then $b(d)$ is the probability that ω_2 *does not cross* ω_1. To say ω_2 *does not cross* ω_1 means, for each horizontal coordinate i shared by vertices (i, j_1) in ω_1 and (i, j_2) in ω_2, we have $j_2 \leq j_1$. In the case $j_2 = j_1$, we say ω_1 and ω_2 *intersect* at $(i, j_1) = (i, j_2)$. Thus we seek all pairs (ω_1, ω_2) such that ω_1 is a path from $(0, 0)$ to (n, d) including the vertex $(n-1, d-1)$, ω_2 is a path from $(0, 0)$ to (n, d') for $1 \leq d' \leq d$, and ω_2 does not cross ω_1. Such a pair (ω_1, ω_2) is illustrated by Figure 2.1. Note that ω_1 and ω_2 may intersect several times. The probability we seek is

$$b(d) = \sum_{n=d}^{\infty}\sum_{d'=1}^{d-1}\sum_{(\omega_1,\omega_2)} P(\omega_1)P(\omega_2),$$

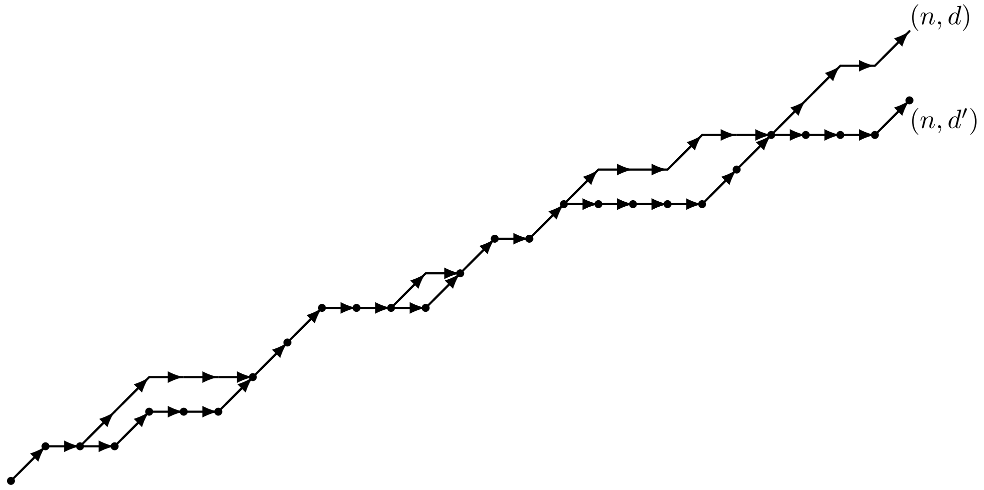
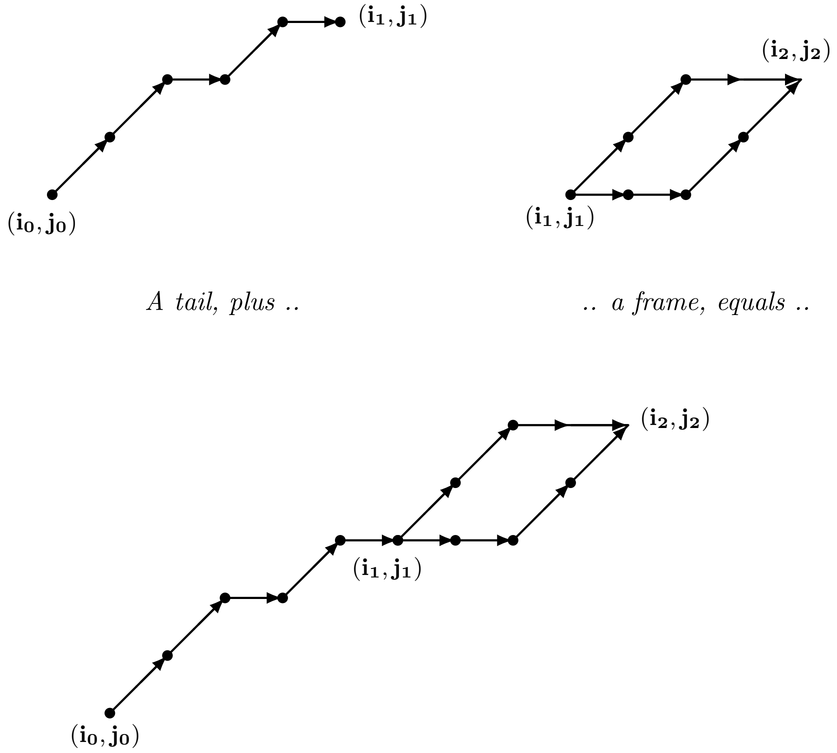


Fig. 2.1. *The winner is never behind.*


$(\mathbf{i_1}, \mathbf{j_1})$
$(\mathbf{i_0}, \mathbf{j_0})$
$(\mathbf{i_2}, \mathbf{j_2})$
$(\mathbf{i_1}, \mathbf{j_1})$
$(\mathbf{i_2}, \mathbf{j_2})$
$(\mathbf{i_1}, \mathbf{j_1})$
$(\mathbf{i_0}, \mathbf{j_0})$

A tail, plus ..

.. a frame, equals ..

Fig. 2.2. *A kite.*

where the innermost sum ranges over all pairs described above.

Look again at Figure 2.1. A pair (ω_1, ω_2) appears to form a chain of flying kites anchored to the ground at $(0, 0)$. The highest kite has two ribbons attached to its tip. Their loose ends are at (n, d) and (n, d').

Each kite consists of a *frame* together with a *tail*. See Figure 2.2. A *frame* from

(i_1, j_1) to (i_2, j_2) consists of a pair of paths from (i_1, j_1), the *lower tip* of the frame, to (i_2, j_2), the *upper tip*, which intersect only at the endpoints. A *tail* from (i_1, j_1) to (i_2, j_2) consists of two identical paths between these endpoints. The *length* of a tail is the number of vertices in the tail minus one, i.e., the number of edges.

A pair (ω_1, ω_2) such that ω_2 does not cross ω_1 forms an alternating sequence of tails and frames, beginning with a tail. Note that tails may have length zero.

The upper tip of the final frame in this sequence is the common endpoint for two paths which intersect only at this common endpoint (these are the "ribbons" described above). One path ends at (n, d), this is the *top ribbon*, and the other ends at (n, d'), the *bottom ribbon*.

Below we compute the probability of a frame from (i_1, j_1) to (i_2, j_2), a tail from (i_1, j_1) to (i_2, j_2), and a pair of ribbons with common initial point (k, d'') and terminal points at (n, d) and (n, d'), respectively.

Let $f_{(i_1,j_1)}^{(i_2,j_2)}(d)$ denote the probability of a frame from (i_1, j_1) to (i_2, j_2). Note for $f_{(i_1,j_1)}^{(i_2,j_2)}(d) \neq 0$, we must have $i_2 \geq i_1 + 2$, $j_2 > j_1$, and $j_2 - j_1 \leq i_2 - i_1 - 1$. Assuming these conditions, we write

$$f_{(i_1,j_1)}^{(i_2,j_2)}(d) = \sum_{(\alpha,\beta)} P(\alpha)P(\beta), \tag{2.27}$$

where (α, β) is a pair of paths from (i_1, j_1) to (i_2, j_2) intersecting only at the endpoints such that β does not cross α. (That is, α forms the upper edge of the frame, and β forms the lower edge.)

We convert the sum above into a determinant using the Gessel–Viennot theorem. Evaluation of the determinant gives

$$f_{(i_1,j_1)}^{(i_2,j_2)}(d) = \frac{j_1 j_2 (d-j_1)!^2}{d^{2(i_2-i_1)}(j_2-j_1)!^2(d-j_2)!^2} \sum_{l,m=j_1}^{j_2} (-1)^{l+m}(lm)^{i_2-i_1-2}(l-j_1)(m-j_2)\binom{j_2-j_1}{m-j_1}\binom{j_2-j_1}{l-j_1}.$$

We compute the probability $t_{(i_1,j_1)}^{(i_2,j_2)}(d)$ of a tail from (i_1, j_1) to (i_2, j_2) in a manner analogous to the computation of $\sum_\alpha P(\alpha)^2$ in section 2.6. We obtain

$$t_{(i_1,j_1)}^{(i_2,j_2)}(d) = \frac{(d-j_1)!^2}{d^{2(i_2-i_1)}(2j_2)!(d-j_2)!^2} \sum_{m=1}^{j_2} (-1)^{j_2-j_1} m^{2(i_2-i_1)}(2m)!\binom{2j_2}{j_2+m}\binom{m+j_1-1}{j_1-m}.$$

Finally we compute the probability $r(d, d', d'', k, n)$ of a pair of ribbons with common initial point (k, d'') and terminal points (n, d') and (n, d). The probability is given by a determinant similar to the one in (2.9). In the present case, we have d'' in place of d_1 and d' in place of d_2. Thus

$$r(d, d', d'', k, n) = \det(d, d', d'', k, n).$$

3. Winning margin. Now we look for the probability distribution of the number of distinct coupons that the second player has collected at the moment the first player completes the collection. Let $g(d,d')$ denote the probability that the second player has collected exactly d' distinct coupons at that moment. The probability that the first player finishes after exactly n trials is $p(n,d)$; see (1.1). The probability that the second player has exactly d' distinct coupons after n trials, given that the first player has just completed the collection at that time, is

$$\binom{d}{d'}\left\{ {n \atop d'} \right\}\frac{d'!}{d^n}$$

if $1 \le d' < d$. Thus for $d' < d$ our distribution g is given by

$$\begin{aligned} g(d,d') &= \sum_{n \ge d'} \frac{d!}{d^n}\left\{ {n-1 \atop d-1} \right\}\binom{d}{d'}\left\{ {n \atop d'} \right\}\frac{d'!}{d^n} \\ &= \frac{d!^2}{(d-d')!}\sum_{n\ge d'}\left\{ {n-1 \atop d-1} \right\}\left\{ {n \atop d'} \right\}\frac{1}{d^{2n}} \\ &= \frac{d!^2}{(d-d')!d^{2d'}}\sum_{r=1}^{d-1} r^{d'-d}\sum_{s=1}^{d'}\frac{A_{d-1,r}A_{d',s}}{1-\frac{rs}{d^2}} - \delta_{d',1}, \end{aligned}$$

by (2.3). A table of the probabilities $\{g(6,j)\}_{j=1}^{5}$ is as follows:

$$.000003, .000793, .018444, .118454, .333986.$$

These do not sum to 1 because the second player might have completed a collection at some time before the first player did.

4. The "double dixie cup problem," of Newman and Shepp, revisited. Here we consider a different generalization of the coupon collector's problem. Let integers $h, d \ge 1$ be fixed. Again we are sampling with replacement from d kinds of coupons, but now T is the epoch at which we have collected at least h copies of each of the d coupons for the first time. (For example, my $h-1$ siblings and I might each want to have our own copy of every one of the available baseball cards.) We study the expectation, the probability generating function, and the asymptotic behavior of the expectation of this generalized problem.

These questions were investigated by Newman and Shepp [4] and the asymptotics were refined by Erdős and Rényi [2]. It is interesting to note that this problem is equivalent to one about the evolution of a random graph. Suppose we fix n vertices, and then we begin to collect from among n kinds of coupons. If we collect a particular sequence, say, $\{c_1, c_2, c_3, \dots\}$, then we add the edges (c_1, c_2),$(c_3, c_4)\dots$. That is, we add an edge each time we choose a new pair of coupons. Our problem about collecting at least h copies of each kind of coupon is thereby equivalent to the question of obtaining a minimum degree of at least h in an evolving random graph.[1] In this section we will not add anything new to the asymptotics of this problem. Instead we claim only a derivation simpler than the original and an explicit generating function, which gives a nice road to the asymptotics. We deal only with generating functions in one variable, whereas in [4] multivariate generating functions were used. We obtain not only the expectation of the time to reach a collection that has at least h copies of each kind of coupon, but also the complete probability distribution of that time.

[1] Our thanks to Ed Bender and to a helpful referee for pointing this out.

For n fixed, consider a sequence of n drawings of coupons that constitutes, for the first time at the nth drawing, a complete collection of at least h copies of each of the d kinds of coupons.

There are d possibilities for the coupon that completes the collection on the nth drawing. There are $\binom{n-1}{h-1}$ ways to choose the set of earlier drawings on which that last coupon type occurred. On the remaining $n-h$ drawings we can define, as usual, an equivalence relation: two drawings i, j are equivalent if the same kind of coupon was drawn at the ith and the jth drawings. The number of such equivalence relations is equal to the number of ordered partitions of a set of $n-h$ elements into $d-1$ classes, each class containing at least h elements. We will denote this latter number by $(d-1)!\left\{{n-h \atop d-1}\right\}_h$, where the $\left\{{n \atop k}\right\}_h$'s count the unordered partitions of an n-set into k classes of at least h elements each.

The number of sequences of n drawings for which we achieve a complete collection for the first time at the nth drawing is therefore

$$d\binom{n-1}{h-1}(d-1)!\left\{{n-h \atop d-1}\right\}_h.$$

Since there are d^n possible drawing sequences of length n, the probability that $T = n$ is

$$p_n = \frac{d!}{d^n}\binom{n-1}{h-1}\left\{{n-h \atop d-1}\right\}_h, \tag{4.1}$$

and the probability generating function is

$$\begin{aligned} P_h(x) \overset{\text{def}}{=} \sum_{n\geq 0} p_n x^n &= \sum_{n\geq 0}\frac{d!}{d^n}\binom{n-1}{h-1}\left\{{n-h \atop d-1}\right\}_h x^n \\ &= d!\binom{xD-1}{h-1}\sum_{n\geq 0}\left\{{n-h \atop d-1}\right\}_h\left(\frac{x}{d}\right)^n, \end{aligned} \tag{4.2}$$

where $D = \partial/\partial x$.

It remains to find the ordinary power series generating function of the $\left\{{n \atop k}\right\}_h$'s. The exponential formula immediately gives us their exponential generating function as

$$\sum_{n\geq 0}\left\{{n \atop k}\right\}_h \frac{x^n}{n!} = \frac{1}{k!}\left(e^x - 1 - x - \cdots - \frac{x^{h-1}}{(h-1)!}\right)^k. \tag{4.3}$$

We can convert this into an ordinary power series generating function by applying the Laplace transform operator

$$\int_0^\infty e^{-sx}\cdots dx$$

to both sides, which yields

$$\sum_{n\geq 0}\left\{{n \atop k}\right\}_h \frac{1}{s^{n+1}} = \frac{1}{k!}\int_0^\infty e^{-sx}\left(e^x - 1 - x - \cdots - \frac{x^{h-1}}{(h-1)!}\right)^k dx,$$

or finally

$$\sum_{n\ge 0}\left\{{n \atop k}\right\}_h t^n = \frac{1}{k!t}\int_0^\infty e^{-x/t}\left(e^x-1-x-\cdots-\frac{x^{h-1}}{(h-1)!}\right)^k dx. \tag{4.4}$$

Now if we substitute (4.4) into (4.2) we obtain the probability generating function of the generalized coupon collector's problem in the form

(4.5)

$$P_h(x)=\frac{1}{d^{h-2}}\int_0^\infty\left\{\binom{xD-1}{h-1}x^{h-1}e^{-td/x}\right\}\left(e^t-1-t-\cdots-\frac{t^{h-1}}{(h-1)!}\right)^{d-1}dt.$$

In the above, $\binom{xD-1}{h-1}$ is the differential operator that is defined by

$$\binom{xD-1}{h-1}f(x)=\frac{1}{(h-1)!}\left(x\frac{d}{dx}-1\right)\left(x\frac{d}{dx}-2\right)\dots\left(x\frac{d}{dx}-h\right)f(x).$$

However, it is easy to establish, by induction on h, the interesting fact that

$$\binom{xD-1}{h-1}x^{h-1}e^{-td/x}=\frac{(td)^{h-1}}{(h-1)!}e^{-td/x}. \tag{4.6}$$

Hence we have proved the following evaluation.

Theorem 4.1. *The probability generating function for the coupon collecting problem in which at least h copies of each coupon are needed is given by*

$$P_h(x)=\frac{d}{(h-1)!}\int_0^\infty t^{h-1}e^{-td/x}\left(e^t-1-t-\cdots-\frac{t^{h-1}}{(h-1)!}\right)^{d-1}dt. \tag{4.7}$$

4.1. Two examples. Let's look at the cases $h=1$, the classical case, and $h=2$, where we want to collect at least two specimens of each of the d kinds of coupons.

If $h=1$, then (4.7) takes the form

$$P_1(x)=d\int_0^\infty e^{-td/x}(e^t-1)^{d-1}dt.$$

If we expand the power of (e^t-1) by the binomial theorem and integrate termwise, we obtain

$$P_1(x)=xd\sum_{j=0}^{d-1}\binom{d-1}{j}\frac{(-1)^{d-1-j}}{d-jx},$$

which is precisely the partial fraction expansion of the classical generating function (2.2).

To see something new, let $h=2$. Then

$$P_2(x)=d\int_0^\infty te^{-td/x}\left(e^t-1-t\right)^{d-1}dt. \tag{4.8}$$

Again, by termwise integration this can be made fairly explicit, but since the most interest attaches to the expectation, let's look at the average number of trials that are needed to collect at least two samples of each of d coupons. This is $P_2'(1)$, which after some simplification takes the form

$$P_2'(1) = d^2 \int_0^\infty \left(\frac{t^2}{e^t - 1 - t} \right) (1 - (1+t)e^{-t})^d dt. \tag{4.9}$$

From this we can go in either of two directions: an exact evaluation or an asymptotic approximation. By termwise integration it is easy to obtain the following exact formula, which is a finite sum, for $\langle T \rangle_2$, the average number of trials needed to collect at least two of each of the d kinds of coupons:

$$\langle T \rangle_2 = d^2 \sum_{m,j} (-1)^m \binom{d-1}{m} \binom{m}{j} \frac{(j+2)!}{(m+1)^{j+3}}. \tag{4.10}$$

For $d = 2, 3, 4, 5$ these are $2, 11/2, 347/36, 12259/864$. To facilitate comparison with the classical ($h = 1$) case, we show below, for $1 \leq d \leq 10$, a table of the expected numbers of trials needed when $h = 1, 2$.

d :	1	2	3	4	5	6	7	8	9	10
$\langle T \rangle_1$:	1.0000	3.0000	5.5000	8.3333	11.417	14.700	18.150	21.743	25.460	29.290
$\langle T \rangle_2$:	2.0000	5.5000	9.6389	14.189	19.041	24.134	29.425	34.885	40.492	46.230

4.2. Asymptotics. Now we investigate the asymptotic behavior of (4.9), for large d, to compare it with the $d \log d$ behavior of the classical case where $h = 1$.

THEOREM 4.2. *If there are d different kinds of coupons, and if at each step we sample one of the d kinds with uniform probability, let $\langle T \rangle_h$ denote the average number of samples that we must take until, for the first time, we have collected at least h specimens of each of the d kinds of coupons. Then for every $h \geq 1$, we have $\langle T \rangle_h \sim d \log d$ $(d \to \infty)$.*

Consider first the case $h = 2$. In (4.9) we make the substitution

$$e^{-u} = 1 - (1+t)e^{-t}, \tag{4.11}$$

where u is a new variable of integration. We then find that

$$P_2'(1) = d^2 \int_0^\infty t(u) e^{-ud} du, \tag{4.12}$$

where $t(u)$ is the inverse function of the substitution (4.11), which is well defined since the right side of (4.11) increases steadily from 0 to 1 as t increases from 0 to ∞.

The main contribution to $P_2'(1)$ comes from values of u near $u = 0$, and when u is near 0 we have

$$t(u) = -\log u + O(\log \log u).$$

Following the arguments in [6, sect. 2.2], we see that $P_2'(1)$ of (4.12) has the same asymptotic behavior as

$$d^2 \int_0^c (-\log u) e^{-ud} du \qquad (0 < c < 1),$$

and in [6] this is shown to be

$$\sim d^2 \cdot \frac{\log d}{d} = d \log d.$$

Now we consider the asymptotic behavior of the expected number of trials for general values of h. From (4.7) we see that this expected number of trials can be written in the form

$$\frac{d^2}{(h-1)!}\int_0^\infty \left\{\frac{t^h}{e^t-1-t-\cdots-\frac{t^{h-1}}{(h-1)!}}\right\}\left\{1-\left(1+t+\frac{t^2}{2}+\cdots+\frac{t^{h-1}}{(h-1)!}\right)e^{-t}\right\}^d dt. \tag{4.13}$$

Again we make the change of variable

$$e^{-u}=1-\left(1+t+\frac{t^2}{2}+\cdots+\frac{t^{h-1}}{(h-1)!}\right)e^{-t} \tag{4.14}$$

in the integral, and it takes the remarkably simple form (compare (4.12))

$$P_h'(1)=d^2\int_0^\infty t(u)e^{-ud}du,$$

where $t(u)$ is the inverse function of the substitution (4.14). Again the main contribution to the integral comes from small values of u, and when u is small and positive we have

$$t(u)=-\log u+(h-1)\log(-\log u)+\cdots.$$

Using the method of section II.2 of [6] once more, we find that

$$\langle T\rangle_h = d\log d+(h-1)d\operatorname{loglog} d(1+o(1)) \qquad (d\to\infty). \tag{4.15}$$

We remark that in the case of $d=200$ coupons, the correct expected number of trials to obtain two of each coupon is 1614 trials, the approximation $d\log d$ is 1175, and the approximation $d\log d+(h-1)d\operatorname{loglog} d$ is 1393, each rounded to the nearest integer.

5. The number of singletons. In view of the asymptotics in the preceding section we realize that at the moment when a coupon collector sequence terminates with a complete collection, "most" coupons will have been collected more than once, and only "a few" will have been collected just once. We call a coupon that has been seen just once a *singleton*. We will now look at the distribution of singletons.

In more detail, let j be the number of singletons in a collecting sequence that terminates successfully at the nth step. We first want the joint distribution $f(n,j)$ of n and j, i.e., the probability that a collecting sequence halts successfully at the nth step and has exactly j singletons at that moment. We claim that

$$f(n,j)=\frac{d!}{d^n}\binom{n-1}{j-1}\left\{ {n-j \atop d-j} \right\}_2. \tag{5.1}$$

Indeed, the last coupon to be collected can be chosen in d ways; the other $j-1$ singleton coupons can be chosen in $\binom{d-1}{j-1}$ ways and can be presented in an ordered sequence in $(j-1)!\binom{d-1}{j-1}$ ways. This ordered sequence can appear among the first $n-1$ trials in $\binom{n-1}{j-1}$ ways, and the remaining $n-j$ trials constitute an ordered partition of $n-j$ elements into $d-j$ classes, no class having fewer than two elements, which can be chosen in $(d-j)!\left\{ {n-j \atop d-j} \right\}_2$ ways. If we multiply these together and divide by d^n, the number of n-sequences, we obtain the result (5.1) claimed above.

Next we compute the probability that a completed collecting sequence contains exactly j singletons, whatever the length of the sequence may be. That is, we find $F(j) = \sum_n f(n,j)$, where f is given by (5.1). We have, after using the generating function (4.4),

$$F(j) = \sum_n \frac{d!}{d^n}\binom{n-1}{j-1}\left\{ {n-j \atop d-j} \right\}_2 \tag{5.2}$$

$$= \frac{d!}{(d-j)!d^j}\left\{\int_0^\infty \left\{ \binom{t\frac{\partial}{\partial t}+j-1}{j-1}\left(\frac{e^{-xt}}{t}\right)\right\}(e^x-1-x)^{d-j}dx\right\}_{t\to 1/d}. \tag{5.3}$$

But using the fact that, analogously to (4.6), we have

$$\binom{t\frac{\partial}{\partial t}+j-1}{j-1}\left(\frac{e^{-xt}}{t}\right) = \frac{x^{j-1}}{(j-1)!t^j}e^{-x/t},$$

we can simplify the expression for $F(j)$ to

$$F(j) = j\binom{d}{j}\int_0^\infty x^{j-1}(e^x-1-x)^{d-j}e^{-xd}dx \qquad (j=1,2,3,\ldots), \tag{5.4}$$

which is the desired distribution of the number of singletons in a successfully terminated coupon collecting sequence.

Now if we multiply by j and sum over j, we'll get the average number of singletons that appear in a completed collection of d coupons. This is, after some termwise integration,

$$\bar{j}(d) = d\sum_m (-1)^m \binom{d-2}{m}\frac{d(m+1)+1}{(m+2)^2(m+1)}.$$

If we expand the summand in partial fractions, viz.

$$\bar{j}(d) = d\sum_m (-1)^m \binom{d-2}{m}\left(\frac{1}{m+1}-\frac{1}{m+2}+\frac{d-1}{(m+2)^2}\right),$$

then each of the three sums indicated can be expressed in closed form, in two cases by using the identity

$$\sum_k (-1)^k\binom{n}{k}\frac{1}{x+k} = \frac{1}{x\binom{x+n}{n}}, \tag{5.5}$$

directly, with $x=1$ and $x=2$, and in the third case by differentiating (5.5) w.r.t. x and using the result with $x=2$. The identity (5.5) is itself certified, after multiplying by the denominator on the right, by the WZ proof certificate $R(n,k) = k(x+k)/((n+1)(k-n-1))$.

What results is that $\bar{j}(d) = H_d$, the dth harmonic number. That is, *the average number of singleton coupons in a completed collection sequence of d coupons is the harmonic number H_d.*

REFERENCES

[1] L. Comtet, *Advanced Combinatorics*, D. Reidel, Dordrecht, The Netherlands, 1974.

[2] P. Erdős and A. Rényi, *On a classical problem of probability theory*, Magyar Tud. Akad. Mat. Kutató Int. Közl., 6 (1961), pp. 215–220 (English. Russian summary).

[3] I. Gessel and G. Viennot, *Binomial determinants, paths, and hook length formulae*, Adv. Math., 58 (1985), pp. 300–321.

[4] D. J. Newman and L. Shepp, *The double dixie cup problem*, Amer. Math. Monthly, 67 (1960), pp. 58–61.

[5] H. S. Wilf, *generatingfunctionology*, 2nd ed., Academic Press, Boston, 1994.

[6] R. Wong, *Asymptotic Approximations of Integrals*, Academic Press, Boston, 1989.

SIAM J. DISCRETE MATH.
Vol. 17, No. 1, pp. 18–49

INTERNET ROUTING AND RELATED TOPOLOGY ISSUES[*]

WALID BEN-AMEUR[†] AND ERIC GOURDIN[‡]

Abstract. In most domains of the Internet network, the traffic demands are routed on a single-path defined as the shortest one according to a set of administrative weights. Most of the time, the values set by the administrator (or the default ones) are such that there are many paths of the same length between the extremities of some demands. However, if the shortest paths are not unique, it might become difficult for an Internet domain administrator to predict and control the traffic flows in the network. Moreover, the sequence order of packets can be changed when many paths are used leading to some end-to-end delays. It is hence an important issue to ensure that each shortest path is unique according to a given set of administrative weights. We show that it is possible to determine a set of small integer weights (smaller than 6 times the radius of the network) such that all links are used and every demand is routed on a unique shortest path. Above and beyond this uniqueness requirement, network administrators wishing to exploit the available resources would like to control the whole routing pattern. The problem they face consists of determining a set of weights enforcing a given routing policy. We formulate this problem using linear programs, and we show how integer weights can be computed by heuristics with guaranteed worst-case performances. Some conditions on the given routing, necessary for the existence of a solution, are derived. Both necessary and sufficient conditions are also provided, together with some other useful properties, in the case of particular graphs such as cycles and cacti.

Key words. Internet network, shortest path routing, graph topology

AMS subject classifications. 05C05, 05C38, 05C85, 65K05, 90B18

DOI. 10.1137/S0895480100377428

1. Introduction.

During the last few years, Internet usage has grown rapidly and there are huge bandwidth requirements for most telecommunication services. Many telecommunication companies are even building their own Internet backbone network designed for large traffic volumes.

The routing of the flows of traffic in an Internet network is completely determined by the choice of a routing protocol and the setting of its parameters. This latter task is devoted to a so-called network administrator or supervisor. The network administrator will usually set or try to set these routing parameters in order to achieve several goals: the first and most obvious one is to optimize the performances of the network. Although this criterion can be interpreted in many different ways, it most often reduces to avoid congestion and therefore to the ability to direct flow where network resources are available. However, this is not the only goal of the network administrator. One of his main duties is indeed to administrate, that is, to know exactly what is going on in the network and to know how to react when something goes wrong. Therefore, the natural trend is to keep things as simple as possible and to avoid the use of too many parameters.

From a pure optimization point of view, there is no doubt that the most efficient way to avoid congestion, for instance, by keeping the maximum load as low as possible, would be to split the traffic freely over all the network. However, it is a maybe not so

[*] Received by the editors August 31, 2000; accepted for publication (in revised form) January 14, 2003; published electronically September 17, 2003.

http://www.siam.org/journals/sidma/17-1/37742.html

[†] Institut National des Télécommunications, 9 rue Charles Fourier, 91011 Evry, France (walid.benameur@int-evry.fr).

[‡] France Telecom R&D, 38–40, rue du Général Leclerc, 92794 Issy–les–Moulineaux, France (eric.gourdin@francetelecom.com).

well-known result that an optimal solution of a multicommodity flow problem with known capacities (at least if obtained by a simplex based method, which is most often the case) will use a mean number of paths per demand between 1 and $1 + \frac{|E|}{|K|}$, where $|K|$ is the number of demands and $|E|$ is the number of edges (see, e.g., [5]). For most real communication networks, the number of edges is about $2n$ (two times the number of vertices) and the number of demands is $n(n-1)/2$. It follows that $\frac{|E|}{|K|}$ is usually very small (from 0.2 for a 20 node network up to 0.05 for an 80 node network) and almost all the demands are hence routed on a single-path. Of course, it is possible to build examples for which the difference between multipath routing and single-path routing is very large (see, e.g., [27, 5]), but this situation seldom occurs in practical instances. Another important result related to routing is given in [13]. It was shown in this paper that any single-source multipath routing can be transformed into a single-path routing with an increase in terms of link loads bounded by the value of the maximum demand. Moreover, when the network is a ring, an optimal multisource multipath routing can be transformed into a single-path routing with an increase in terms of link loads bounded by 3/2 times the maximum demand [35].

From a practical point of view, the optimal routing pattern obtained as a result of an optimization process must be implemented in practice and therefore must be compatible with a given Internet routing protocol. Most of the backbone Internet networks still use some classical Internet routing protocols such as open shortest path first (OSPF), RIP, or IS-IS [23, 24, 30, 32, 34] to route the demands. These protocols are based on shortest path routing. Path length is defined as the sum of weights associated with the links of the path. These weights, often called administrative weights, are managed by the network administrator. Ideally, the administrator would like to modify some of the current routing paths in order to better exploit the available resources. In practice, the administrator can only manipulate link weights (and not complete end-to-end routing paths) and the goal is therefore difficult to achieve. That is why the link weights are very often set to some default values such as 1, or the inverse of the link capacity. More advanced versions of OSPF allow indeed to split the traffic on several shortest paths (load balancing) but the practical implementation of these mechanisms is complicated. For instance, in the ECMP (equal cost multipath) mechanism, the load balancing can only be even (same proportion of traffic on each path) and it can only be realized on shortest paths (i.e., equal length paths which length is also the shortest). In order to set up efficiently such a mechanism, the administrator must set the weights in order that the all paths he has chosen be shortest path. Besides, in the optimization process, he must take into account the fact that the traffic can only be evenly splitted. Such an optimization problem is very difficult to solve (NP-hard) [17, 19]. There are other difficulties implied by the traffic splitting. Indeed, if the splitting is done on a packet per packet basis, the resequencing of packets must be handled at some point, in order for the network to support in-order packet delivery. On the other hand, if many paths are used for the same commodity, the size of the routing tables will also increase significantly. Even if the Internet routing devices seem to become more and more efficient, the congestion of modern communication networks is still very often due to congestions occurring in the nodes rather than on the links. In fact, the transmission rates based on optical technologies (wavelength division multiplexing) increased considerably during the last few years, as compared to the evolution of node processors speed. Finally, the new routing paradigm based on the explicit definition of end-to-end tunnels (multiprotocol label switching (MPLS), tag switching, . . .) seem very promising in the sense that they allow much more flexibility in the management of the routing pattern. However, up to now, there does not seem to exist a consensus on the way to use these new protocols, which are nonetheless still based to some extent on shortest path routing. Indeed, the huge number of control parameters offered in MPLS (ways to define the LSPs, constrained based routing, use of colored edges, ...) is rather introducing a new complexity than easing the day-to-day network management problems. In this context, it is hence very reasonable to consider that some network administrators will prefer to rely on simple routing mechanisms based on existing and well-known protocols, which, if wisely used, can still achieve very good performances.

In this paper, we consider problems in which the demands are routed on *unique* shortest paths. This uniqueness requirement allows the network administrators to avoid all the technical difficulties related with the management of load balancing. The family of problems addressed in this paper are the ones a network administrator might encounter, such as trying not to waste available resources and, at the same time, to impose some predefined routing strategy. For instance, one result that is difficult to achieve consists of determining weights such that each routing path is the *unique* shortest path according to these weights.

Several results related to this domain are already available in the literature. A general multipath routing was also studied in [37, 5]. An optimal multipath routing is computed by linear programming, and weights are deduced by duality in [37]. A more general model computing weights for any multipath routing and integrating any set of practical linear constraints related to weights is given in [5]. It is also shown in [5] how a unique shortest path routing can be transformed into an optimal multipath routing by adding a minimal number of paths (MPLS tunnels). Several papers propose methods to modify the link weights when the network state changes. The weights can depend on link loads, physical link lengths, costs, service classes, etc. [11, 24, 26, 34, 36, 40]. Some heuristics were proposed in [17, 19] to compute weights minimizing, in a certain sense, the network load. The routing here is based on the equal cost splitting capability of OSPF: the traffic is evenly splitted between equal cost paths. The problem of finding an optimal even split of traffic was shown to be difficult [17, 19]. Single-path routing, without weight consideration, has also been studied by many authors. Some solution methods based on cutting plane algorithms are presented in [1, 20, 33]. Approximation algorithms are given in [13].

A comparison between different routing strategies in terms of congestion is done in [27]. A capacitated version of the problems addressed in this paper, where one has to determine simultaneously the weights on the links and the capacities required to route all traffic demands, has already been studied [3, 4]. In these papers, heuristics, meta-heuristics, and even exact methods have been proposed to solve the problem, and the resulting algorithms have been applied on some real network designs. In the papers cited above, the weight of each edge is supposed to belong to a finite and small set of real values. The meta-heuristic moves consist in changing the weights and computing the routing and the capacities that have to be installed on links. The exact method is based on a cutting plane algorithm that considers many families of valid inequalities. Finally, the topology of many kinds of survivable networks, including Internet networks, have been studied in [2]. More general models for Internet topology were proposed in [9, 38, 15].

This paper is devoted to *uncapacitated* problems, which involve determining an optimal set of administrative weights to achieve some predefined requirements on the usage of the network resources or on the paths used to route the demands. More precisely, two problems will be addressed in this paper. In the first one, in order to conform to some Internet routing protocols, the weights on the links must be small integer values. In this problem, we assume that there is a traffic demand for each pair of nodes in the network. The aim is to compute link weights such that

1. the weights are integers and are as small as possible;
2. the shortest path between each pair of nodes is unique;
3. all the network links are used.

We show that it is possible to find weights satisfying these constraints, which are strictly lower than 6 times the radius of the graph associated with the network.

The second problem is defined as follows: given a set of predefined routing paths between some node pairs, we wish to compute weights which are compatible with this set of routing paths. In other words, the computed weights must be such that the unique shortest path for each demand is, indeed, the one chosen in advance. The given set of routing paths must satisfy some conditions to guarantee the existence of a compatible set of weights. Three necessary conditions are provided. In the general case, linear mathematical programming formulations of the problem are proposed. We also show how integer weights can be easily computed using a polynomial heuristic with worst-case guaranteed performance. In the case of some particular graphs (cycle, cactus, clique,...), simple necessary and sufficient conditions for existence of compatible weights are given, and methods to explicitly derive these weights are presented.

The paper is organized as follows: the first problem is addressed in section 2, and section 3 is devoted to the second problem. Computational results are discussed in section 4, and conclusions and perspectives are exposed in section 5.

2. How to get small integer weights and use all links.

2.1. Definitions and notation. Let $G = (V, E)$ denote an undirected graph associated with the Internet network considered, where V is the set of nodes (or vertices) and E the set of links (or edges). The number of nodes is denoted n. The graph G is assumed to be connected. Let R be the radius of G and C the set of central vertices of G.

Recall that a central vertex v is such that the greatest distance from v to any other vertex, called eccentricity of v, is the lowest one. The eccentricity of a central vertex is exactly equal to the radius of the graph [6, 12]. Note that the word “distance” is used to express the number of edges on a path, whereas the word “length” represents the sum of weights along the path. The formal definitions of these terms will be given below.

Assume one particular central vertex has been chosen. Let $T_0 = (V, E_0)$ be a rooted spanning tree of G obtained by a breadth-first search starting from that central vertex. Assume the central vertex is indexed by 0. The remaining vertices of G are then ordered from 1 to $n-1$ according to their marking order during the breadth-first building process of T_0. In what follows, the *order* of a vertex refers to its order in this building sequence. Figure 1 shows a tree T_0 where the edges belonging to the tree are plotted in continuous lines and the other edges of G are plotted in dotted lines. The order of the vertices are displayed in the figure.

Definition 2.1 (levels). *We say that a vertex v is located at level i if the distance between v and 0 is equal to i. The level of a vertex i is denoted level(i).*

The levels are illustrated at the left of the figure. As 0 is a central vertex, the number of levels is equal to the radius R. The unique path linking two vertices a and b in the rooted tree T_0 is denoted $P_0(a, b)$.

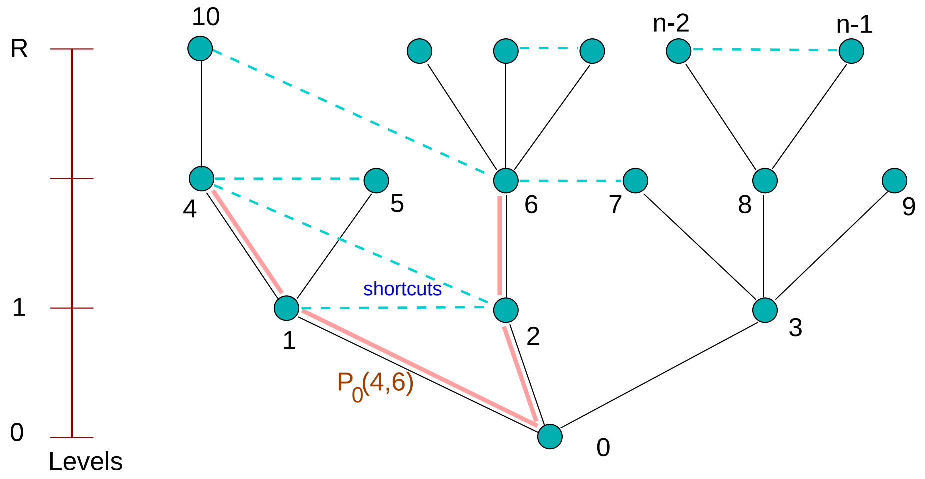


FIG. 1. *Tree T_0 obtained by breadth-first search.*

DEFINITION 2.2 (monotonous path). *Given any vertices a and b of V, the path $P_0(a,b)$ is said to be monotonous if it does not contain two vertices located on the same level.*

This is equivalent to saying that $P_0(a,0) \subset P_0(b,0)$ or $P_0(b,0) \subset P_0(a,0)$.

DEFINITION 2.3 (turning path). *A nonmonotonous path is called a turning path.*

Thus, a turning path $P_0(a,b)$ is made of two monotonous paths $P_0(a,d)$ and $P_0(b,d)$ where d is the vertex of $P_0(a,b)$ having the lowest level. This is due to the fact that, for any vertex v whose level is $i \geq 1$, there is exactly one edge of E_0 linking v to the set of vertices whose level is $i-1$.

DEFINITION 2.4 (turning vertex). *The vertex of lowest level in a turning path is called the* turning vertex*.*

DEFINITION 2.5 (distance on a path). *The number of edges of any path $P(a,b)$ in G is called the distance between the vertices a and b and is denoted $\delta(P(a,b))$. Since a path $P_0(a,b)$ in the tree T_0 is unique, its distance is denoted $\delta_0(a,b)$.*

DEFINITION 2.6 (length of a path). *Assume a weight w_e is associated with each edge $e \in E$. The length of a path $P(a,b)$, denoted $\ell(P(a,b))$, is the sum of the weights on its edges:*

$$\ell(a,b) = \sum_{e \in P(a,b)} w_e.$$

Again, since a path $P_0(a,b)$ is unique and there is no ambiguity in its definition, its length is denoted $\ell_0(a,b)$. Note that if all weights are equal to 1, the distance and the length of a path are equal.

DEFINITION 2.7 (shortcuts). *An edge of $E \setminus E_0$ linking two vertices of $P_0(a,b)$ is called a shortcut of $P_0(a,b)$. The set of all shortcuts for a path $P_0(a,b)$ is denoted $S_0(a,b)$. The number of shortcuts is denoted $k_0(a,b) = |S_0(a,b)|$. As will be established, if $k_0(a,b) > 0$, there is no ambiguity in defining a "highest" shortcut as the shortcut with end-points at the highest level.*

The main result of this section consists of deriving a set of small integer weights satisfying some particular constraints. This result is based on properties of breadth-first search trees.

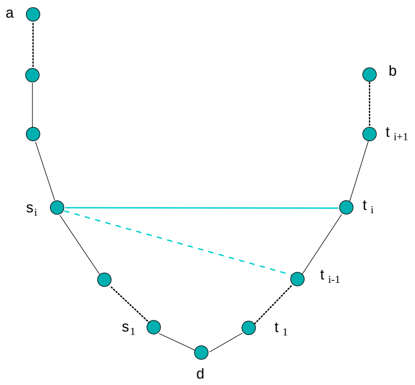


Fig. 2. *A turning path $P_0(a, b)$.*

2.2. Properties of breadth-first search trees.

Lemma 2.8. *If a and b are two adjacent vertices of G such that $ab \in E \setminus E_0$, then the levels of a and b are equal or consecutive.*

Proof. Without loss of generality, suppose that the level of b is not higher than the level of a. As a and b are adjacent, any path from n_0 to b can be completed by the edge ab to obtain a path from n_0 to a. This clearly means that the level of a is at most equal to 1 plus the level of b. □

Lemma 2.9. *Let a and b be two vertices of G. For any vertex c of $P_0(a, b)$, there are at most 2 edges in $G \setminus T_0$ linking c to the other vertices of $P_0(a, b)$.*

Proof. If $P_0(a, b)$ is monotonous, then we can deduce by Lemma 2.8 that there is no edge in $G \setminus T_0$ linking c to the other vertices of $P_0(a, b)$. Let us now assume that $P_0(a, b)$ is a turning path, and let d be the turning vertex of $P_0(a, b)$. We denote $s_0, s_1, \ldots, s_p$ the successive vertices of the path $P_0(a, d)$ starting from d (i.e., $d = s_0$ and $a = s_p$). Similarly, we denote $t_0, t_1, \ldots, t_q$ the successive vertices of the path $P_0(b, d)$ starting from d (see Figure 2). Obviously, for all $i = 0, \ldots, \min\{p, q\}$, the two vertices s_i and t_i are at the same level. Without loss of generality, let us suppose that the order of s_1 is smaller than that of t_1 (i.e., it has been encountered first in the breadth-first process). Using a simple induction, we can deduce that for any two vertices s_i and t_i, the order of s_i is smaller than that of t_i.

Consider a vertex c of the path $P_0(a, b)$. First, we assume that c belongs to the monotonous subpath $P_0(a, d)$. According to our notation, vertex c corresponds to a certain s_i. As $P_0(a, d)$ is monotonous, there is no edge of $G \setminus T_0$ linking c to the other vertices of $P_0(a, d)$. By Lemma 2.8, $c = s_i$ can only be adjacent to vertices on the levels $i - 1$, i, and $i + 1$. Assume that $q \geq i + 1$ (otherwise, the result is obtained), i.e., the path $P_0(b, d)$ includes all vertices at least up to t_{i+1}. Hence the vertex $c = s_i$ is at most linked to t_{i-1}, t_i,s and t_{i+1}. Suppose that s_i is adjacent to t_{i+1}. As the order of s_i is lower than the order of t_i, the tree T_0 should contain the edge $s_i t_{i+1}$ and not $t_i t_{i+1}$. Consequently, c is not adjacent to t_{i+1}. Thus, $c = s_i$ is at most linked to $t_{i-1}$and t_i.

The case where $c \in P_0(b, d)$ can be treated in the same way. □

Lemma 2.10. *Consider a turning path $P_0(a, b)$ between two vertices a and b of G. Let d be the turning vertex of $P_0(a, b)$. Denote by s_i (resp., t_i) the ith vertex in the subpath $P_0(a, d)$ (resp., $P_0(b, d)$ starting from vertex d). For all i such that the vertices s_{i+1} and t_{i+1} exist, the graph G cannot contain at the same time the edges*

s_it_{i+1} and t_is_{i+1}.

Proof. Suppose that the edge t_is_{i+1} exists. By definition of T_0, this means that the order of s_i is lower than the order of t_i. In this case, it is clear that we cannot have an edge s_it_{i+1} (again by definition of T_0). □

LEMMA 2.11. *For all a and b in V such that $a \neq b$, we have $k_0(a,b) \leq \delta_0(a,b)-1$.*

Proof. If $P_0(a,b)$ is monotonous, then $k_0(a,b) = 0$ and the inequality is satisfied. Let us focus on the case where $P_0(a,b)$ is a turning path. Let d be the turning vertex of $P_0(a,b)$. There is no edge in $G \setminus T_0$ linking any two vertices of $P_0(a,d)$ or any two vertices of $P_0(b,d)$ since these paths are monotonous. Thus, $k_0(a,b)$ is given by the number of edges in $G \setminus T_0$ between the two subpaths, excluding the turning vertex d.

First, we assume that $\delta_0(a,d) < \delta_0(b,d)$. Applying Lemma 2.9 to all vertices of $P_0(a,d)\backslash\{d\}$ leads to $k_0(a,b) \leq 2 \times \delta_0(a,d)$. But $2 \times \delta_0(a,d) < \delta_0(a,d) + \delta_0(b,d) = \delta_0(a,b)$, which means that $k_0(a,b) \leq \delta_0(a,b)-1$. By symmetry, the result is also valid when $\delta_0(a,d) > \delta_0(b,d)$.

Let us treat the case $\delta_0(a,d) = \delta_0(b,d)$. The vertices a and b have the same level. It is easy to see that the one having the highest order cannot be linked to more than 1 vertex of $P_0(a,b)$. Consequently, $k_0(a,b) \leq 2 \times \delta_0(b,d) - 1 = 2 \times \delta_0(a,d) - 1 = \delta_0(a,b) - 1$. Thus, the inequality of the lemma is still valid. □

DEFINITION 2.12. *Let a, b, and c be three vertices of G such that $c \in P_0(a,b)$. We define $k_0(a,c,b)$ as the number of edges of $G \setminus T_0$ linking the vertices of $P_0(a,c)\backslash\{c\}$ and the vertices of $P_0(a,b)$.*

Note that any edge linking c to a vertex of $P_0(a,c)$ is taken into account in the definition of $k_0(a,c,b)$. On the other hand, the edges linking c to any vertex of $P_0(c,b)$ are not considered.

LEMMA 2.13. *If a, b, and c are three vertices of G such that $c \in P_0(a,b)$, then $k_0(a,c,b) \leq 2 \times \delta_0(a,c)$.*

Proof. This is shown by applying Lemma 2.9 to all the vertices of $P_0(a,c)$. □

LEMMA 2.14. *If a, b, and c are three vertices of G such that $c \in P_0(a,b)$, then $k_0(a,b) = k_0(a,c) + k_0(b,c,a)$ and $k_0(a,c) + k_0(c,b) \leq k_0(a,b)$.*

Proof. Any edge of $G \setminus T_0$ between two vertices of $P_0(a,b)$ is linking either two vertices of $P_0(a,c)$ or a vertex of $P_0(b,c)\backslash\{c\}$ to any vertex of $P_0(a,b)$. This immediately gives the equality $k_0(a,b) = k_0(a,c) + k_0(b,c,a)$. Combining this equality with the inequality $k_0(b,c,a) \geq k_0(b,c)$ leads to $k_0(a,c) + k_0(c,b) \leq k_0(a,b)$. □

Any path between two vertices a and b in the tree T_0 visiting n vertices $v_1, \dots, v_n$ is obviously at least as long as the direct path $P_0(a,b)$. This result is formalized in the two following propositions.

PROPOSITION 2.15. *Consider three vertices a, b, and c in T_0.*

(i) *Vertex c belongs to the path $P_0(a,b)$ if and only if $\delta_0(a,c) + \delta_0(c,b) = \delta_0(a,b)$.*

(ii) *Otherwise, $\delta_0(a,c) + \delta_0(c,b) - \delta_0(a,b) = 2\ \delta_0(i,c)$, where i is the last vertex of $P_0(a,b)$ also belonging to $P_0(a,c)$ (see Figure 3).*

Proof. The first assertion is obvious. In the case where c does not belong to the path $P_0(a,b)$, then there is a vertex i where the path $P_0(a,c)$ departs from the path $P_0(a,b)$. (In other words, i is the last vertex common to the two paths.) Hence $P_0(a,c) = P_0(a,v) \cup P_0(v,c)$. Similarly, since T_0 is a tree, we have $P_0(c,b) = P_0(c,v) \cup P_0(v,b)$. When counting all the edges from a to b going through c, the edges of $P_0(v,c)$ are counted twice:

$$\delta_0(a,c) + \delta_0(c,b) = \delta_0(a,i) + \delta_0(i,c) + \delta_0(c,i) + \delta_0(i,b) = \delta_0(a,b) + 2\ \delta_0(i,c). \qquad □$$

Note that in the second case, we could have simply stated that the quantity

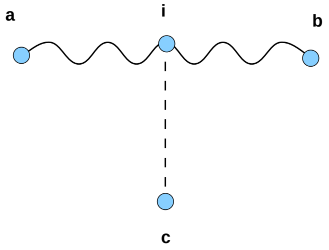


FIG. 3. *A path departing from $P_0(a,b)$.*

$\delta_0(a,c)+\delta_0(c,b)-\delta_0(a,b)$ is even, i.e., equal to $2p$ with a certain positive integer p. This result easily generalizes to n intermediate vertices.

COROLLARY 2.16. *Consider $n+2$ vertices of V, $a=v_0,v_1,\dots,v_n,v_{n+1}=b$ (not necessarily all different).*

(i) *All the vertices $v_1,v_2,\dots,v_q$ belong to the path $P_0(a,b)$ and are encountered in that precise order when going from a to b if and only if $\sum_{i=0}^{n}\delta_0(v_i,v_{i+1})=\delta_0(a,b)$.*

(ii) *Otherwise, there is a positive integer p such that $\sum_{i=0}^{n}\delta_0(v_i,v_{i+1})-\delta_0(a,b)=2\,p$.*

Proof. This result is easily obtained by induction using the previous proposition. □

2.3. Deriving admissible weights. In the problem considered here, it is assumed that the network topology is given (for instance, as a result of an optimization process, or simply to reflect an existing network). We assume that a traffic demand is defined between each pair of nodes. Although this seems rather restrictive, such a requirement is always satisfied in practice on backbone networks, where each demand is in fact the aggregation of a huge number of individual traffic demands. Indeed, the problems considered in this paper are essentially relevant for such backbone networks. The problem then consists of determining a set of link weights satisfying some constraints.

As explained in the introduction, for consistency and manageability reasons, the shortest path between each pair of nodes must be unique. Second, each network link must belong to at least one shortest path. In other terms, all the network links are necessarily used to carry traffic. This requirement avoids leaving some resources unused. Additionally, it means that the traffic must be distributed over the all network, which is highly advisable for reliability reasons. Finally, we want the computed weights to be integers and be as small as possible. This is simply for technical reasons, since many routing protocols such as RIP, IS-IS, or even PNNI in the context of asynchronous transfer mode (ATM) networks [18] require the weights to have bounded integer values ([0,15] for RIP, [0,63] for IS-IS, and [0,65535] for PNNI).

We now focus on two of the main contributions of the paper, namely, the construction scheme to derive a set of integer weights satisfying all the constraints such that all weights are strictly lower than 6 times the radius of the graph and the theorem on which this construction scheme is based.

Construction scheme.

1. Compute a central vertex v (indexed by 0) of the graph G. (Recall that a central vertex is a vertex such that the maximum distance from it to any other vertex is minimal.)
2. Using a breadth-first search starting from v, build a tree T_0.

3. Associate a weight $w_0 \geq 3$ with each edge of T_0.
4. Associate a weight $w_e = w_0 \times \delta_0(i,j) - k_0(i,j)$ with each edge $e = ij \in E \setminus E_0$.

Using this polynomial construction scheme and taking $w_0 = 3$, we obtain a weighted graph such that all desired properties are satisfied.

Theorem 2.17. *Let G be a connected graph and R its radius. The weights associated with each edge according to the above construction scheme are such that*

- the weight of every link is a positive integer $\leq 6R - 1$,

- every link of G belongs to at least one shortest path in the sense of these weights,

- there is exactly one shortest path between any pair of vertices.

The proof of Theorem 2.17 is rather technical. It is provided at the end of the section as the natural consequence of a series of preliminary results.

Note that the value of $6R - 1$ is often low, even for large-size telecommunication networks. We also know (see [7], for example) that if we consider random graphs where an edge exists with a constant probability p, then almost all the graphs have a radius equal to 2. Thus, Theorem 2.17 gives us a set of weights which satisfy the Internet routing constraints and which are lower than 11 in almost all cases. This is valid for random graphs, but if we consider any deterministic graph, we can use the fact that the radius is lower than half the number of vertices. This leads to the obvious corollary.

Corollary 2.18. *For any connected graph G, there exist a set of weights which satisfy Internet routing constraints (those of Theorem* 2.17*) and which are lower than* $3N - 1$.

From now on, we assume that G is a weighted graph and that the weights associated with the edges are defined according to the above construction scheme.

Proposition 2.19. *Consider three vertices a, b, and c in T_0. We then have*

$$\ell_0(a,c) + \ell_0(c,b) - \ell_0(a,b) \geq k_0(a,c) + k_0(c,b) - k_0(a,b).$$

Moreover, if $c \notin P_0(a,b)$, then

$$\ell_0(a,c) + \ell_0(c,b) - \ell_0(a,b) \geq k_0(a,c) + k_0(c,b) - k_0(a,b) + 2.$$

Proof. If $c \in P_0(a,b)$, then the left-hand side is equal to zero, and according to the second part of Lemma 2.14, $k_0(a,c) + k_0(c,b) - k_0(a,b)$ is nonpositive. Thus, in this case, the result holds. Assume now that $c \notin P_0(a,b)$. This means that there is a vertex $i \in P_0(a,b)$ at which the path $P_0(a,c)$ departs from the path $P_0(a,b)$ (see Figure 3). Applying the first part of Lemma 2.14, we derive the following expression:

$$k_0(a,c) + k_0(c,b) - k_0(a,b) = k_0(a,i) + k_0(c,i,a) + k_0(b,i) + k_0(c,i,b) - k_0(a,b).$$

On one hand, according to Lemma 2.14, we have $k_0(a,i) + k_0(b,i) \leq k_0(a,b)$. On the other hand, according to Lemma 2.13, we have $k_0(c,i,a) \leq 2\ \delta_0(c,i)$ and $k_0(c,i,b) \leq 2\ \delta_0(c,i)$. We obtain

$$k_0(a,c) + k_0(c,b) - k_0(a,b) \leq 4\ \delta_0(c,i).$$

Finally, according to Proposition 2.15, we have

$$\ell_0(a,c) + \ell_0(c,b) - \ell_0(a,b) = 2\ w_0\ \delta_0(c,i) \geq 6\ \delta_0(c,i) \geq 4\delta_0(c,i) + 2.$$

The last inequality is valid since $\delta_0(c,i) \geq 1$. Hence, in the case where $c \notin P_0(a,b)$, the second part of the proposition is established and the first part follows. □

Again, this result can easily be generalized by induction.

COROLLARY 2.20. *Consider $n+2$ vertices of V, $a = v_0, v_1, \dots, v_n, v_{n+1} = b$ (not necessarily all different). We then have*

$$\sum_{i=0}^{n} \ell_0(v_i, v_{i+1}) - \ell_0(a,b) \geq \sum_{i=0}^{n} k_0(v_i, v_{i+1}) - k_0(a,b).$$

Moreover, if one vertex v_i does not belong to $P_0(a,b)$, or if the vertices $v_1, \dots, v_n$ are not encountered in this order when going from a to b, then

$$\sum_{i=0}^{n} \ell_0(v_i, v_{i+1}) - \ell_0(a,b) \geq \sum_{i=0}^{n} k_0(v_i, v_{i+1}) - k_0(a,b) + 2.$$

Proof. The first part of the corollary is easy to establish by induction on the number of vertices, using arguments similar to the proof of the second part.

We now provide the proof for the second part of the corollary. First, observe that, according to Proposition 2.19, the result is true for $n = 1$. Assume the result is true for any set of at most $n+1$ vertices. Consider $n+2$ vertices of V, $a = v_0, v_1, \dots, v_n, v_{n+1} = b$, and assume the assertion "all vertices $v_1, \dots, v_n$ belong to the path $P_0(a,b)$ and are ranked in that same order along the path" is false. This is equivalent to saying that either one or both the following two assertions are also false:

(i) "all vertices $v_1, \dots, v_{n-1}$ belong to the path $P_0(a, v_n)$ and are encountered in this order when going from a to v_n";

(ii) "v_n belongs to $P_0(a,b)$".

We will now compute the quantity

$$\Delta = \sum_{i=0}^{n} \ell_0(v_i, v_{i+1}) - \ell_0(a,b) = \sum_{i=0}^{n-1} \ell_0(v_i, v_{i+1}) + \ell_0(v_n, b) - \ell_0(a,b)$$

in both cases.

Assume assertion (i) is wrong. Then, according to the induction hypothesis, we have

$$\sum_{i=0}^{n-1} \ell_0(v_i, v_{i+1}) - \ell_0(a, v_n) \geq \sum_{i=0}^{n-1} k_0(v_i, v_{i+1}) - k_0(a, v_n) + 2.$$

It follows that

$$\Delta \geq \ell_0(a, v_n) + \ell_0(v_n, b) - \ell_0(a,b) + \sum_{i=0}^{n-1} k_0(v_i, v_{i+1}) - k_0(a, v_n) + 2.$$

Applying the first inequality of Proposition 2.19, the result follows immediately.

Assuming assertion (ii) is wrong, we apply the first inequality of the corollary to obtain

$$\Delta \geq \ell_0(a, v_n) + \ell_0(v_n, b) - \ell_0(a,b) + \sum_{i=0}^{n-1} k_0(v_i, v_{i+1}) - k_0(a, v_n).$$

As v_n does not belong to $P_0(a,b)$, we can now apply the second inequality of Proposition 2.19, and again, the result follows. □

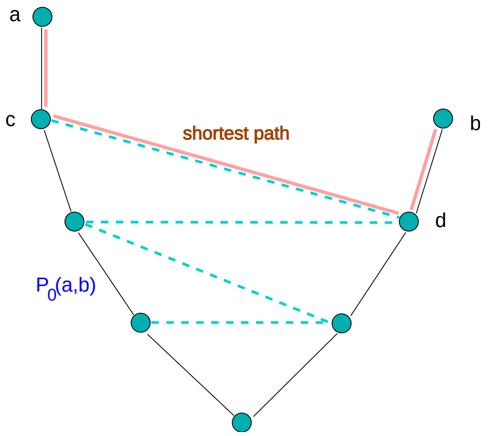


FIG. 4. *Unique shortest path between a and b.*

In order to derive the main result, we now characterize the shortest paths in graph G, according to the weights w_e.

LEMMA 2.21. *For any pair of vertices a and b, if path $P_0(a, b)$ is monotonous, then it is the unique shortest path between a and b.*

Proof. First observe that, according to Lemma 2.8, the edges of E connect only vertices at consecutive levels or at the same level. Hence any path from a to b must at least cross $|level(a)-level(b)| = \delta_0(a, b)$ levels. As all the weights are nonnegative, any path using an edge outside the bounds defined by $level(a)$ and $level(b)$ is longer than $P_0(a, b)$. Consider an edge $e = ij \in E \setminus E_0$. According to Lemma 2.11, we have $k_0(i, j) < w_0\ k_0(i, j) \leq w_0\ \delta_0(i, j) - w_0$, and hence

$$w_e = w_0\ \delta_0(i, j) - k_0(i, j) > w_0.$$

It follows that any path between a and b using such an edge is necessarily longer than $P_0(a, b)$. ∎

COROLLARY 2.22. *Consider the family of paths between the vertices a and b using edges of T_0 and any shortcut of $P_0(a, b)$. The unique shortest path among all these paths is*

(i) $P_0(a, b)$ *if* $k_0(a, b) = 0$,

(ii) $P_0(a, c) \cup cd \cup P_0(d, b)$, *where cd is the highest shortcut for $P_0(a, b)$, if* $k_0(a, b) > 0$.

Proof. First note that, if $k_0(a, b) = 0$, then $P_0(a, b)$ is the unique path between a and b among the ones considered. Assume now that $k_0(a, b) > 0$. Consider a shortcut $cd \in S_0(a, b)$ such that $P_0(a, c)$ and $P_0(d, b)$ are monotonous (see Figure 4). According to the previous lemma, the unique shortest paths between a and c on one hand and between d and b, on the other hand are, respectively, $P_0(a, c)$ and $P_0(d, b)$. Hence the length of the shortest among all paths using the shortcut cd is

$$w_0\ \delta_0(a, c) + (w_0\ \delta_0(c, d) - k_0(c, d)) + w_0\ \delta_0(d, b) = w_0\ \delta_0(a, b) - k_0(c, d).$$

The minimum of all such values is achieved using the highest possible value of $k_0(c, d)$ for all shortcuts of $P_0(a, b)$, that is, $k_0(a, b)$. This value is only achieved by the path using the highest shortcut. ∎

PROPOSITION 2.23. *A shortest path in the whole graph G between two vertices a and b cannot contain an edge $e = cd \notin T_0$ such that c or d (or both) do not belong to $P_0(a,b)$.*

Proof. We will establish this result by induction on the number k of edges of any path $P(a,b)$ using such an edge $e = cd$. The result is obviously true when $k = 1$. Assume it is true for all paths up to $k-1$ edges. Consider a shortest path $P(a,b)$ having k edges. Assume it contains an edge $e = cd$ as described in the proposition. Without loss of generality, we can assume that the vertex c is encountered before d in the path $P(a,b)$. Both the subpaths of $P(a,b)$ going from a to c and from d to b have fewer than k edges. Applying the induction hypothesis, these subpaths can only be shortest if they do not contain edges other than shortcuts and edges of the tree T_0. Hence, according to Corollary 2.22, the length of these subpaths is, respectively, $w_0\delta_0(a,c) - k_0(a,c)$ and $w_0\delta_0(d,b) - k_0(d,b)$. (Note that, for instance, if there is no shortcut between a and c, then we have $k_0(a,c) = 0$ and the subpath is simply $P_0(a,c)$.) Consequently, the length of the shortest path $P(a,b)$ is

$$\ell_0(a,c) + \ell_0(c,d) + \ell_0(d,b) - k_0(a,c) - k_0(c,d) - k_0(d,b) \geq \ell_0(a,b) - k_0(a,b) + 2.$$

The inequality is obtained thanks to Corollary 2.20. Besides, $\ell_0(a,b) - k_0(a,b)$ is the length of a valid path from a to b, which is shorter than $P(a,b)$. This shows that the shortest path $P(a,b)$ cannot contain an edge $e = cd$ which is not a shortcut of $P_0(a,b)$. □

We have proved that a shortest path can contain only shortcuts or edges of the tree T_0. The result of Corollary 2.22 can hence be extended to all paths from a to b. It follows that there is exactly one shortest path between any pair of vertices. Besides, each edge $ab \in E_0$ is a shortest path $P_0(a,b)$. Finally, if we take $w_0 = 3$, then all weights are obviously positive and strictly lower than $6R$, and proof of Theorem 2.17 is hence established.

Observe that the shortest paths defined in Corollary 2.22 (see Figure 4) and based on the weights of the construction scheme do not depend on the exact value of w_0. They are the same for any value of w_0 which is higher than 3. In other terms, the vector whose components are the $\delta_0(x,y)$ is a sort of neutral element for the routing paths of Theorem 2.17. In fact, we will see in the next section that, for a given set of routing paths, the set of weights compatible with them is a polyhedron.

3. Achieving weights compatible with a set of routing paths. In this section, we are concerned with the problem where all routing paths are given, but the weights to put on the edges such that each of the given routing paths is a unique shortest path are unknown. Moreover, it is not even known if such weights exist.

Such a problem arises naturally in the context where the routing paths are provided by a dimensioning process (such as, for instance, the one presented in [4]) or in many practical situations where some paths are imposed due to technical reasons (such as transmission delays, link loads, costs, etc.). If we assume that these given routing paths have to be realized by means of a classical Internet routing protocol, then each given path has to be a shortest path according to a set of edge weights. In order to avoid ambiguity in the choice of the path effectively used, we also require that each shortest path be unique. We use the term "compatibility" to characterize a set of edge weights realizing a set of given routing paths.

DEFINITION 3.1 (compatibility). *A set of edge weights is said to be compatible with a set of routing paths if each routing path is the unique shortest path between its extremities according to the set of weights.*

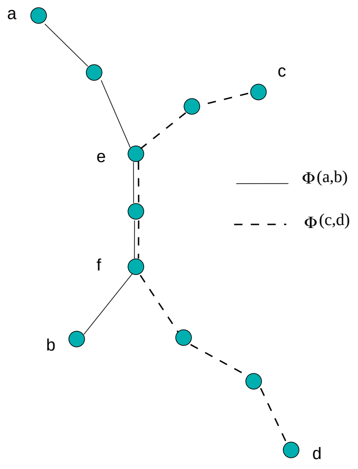


Fig. 5. *Suboptimality condition.*

We are thus concerned with the problem of determining a set of weights compatible with a set of routing paths, if such a set exists, or confirming that no such set exists. A similar problem was studied in [8] in the context of seismography. The uniqueness constraint on the shortest path was not considered in [8], and the problem was formulated as a quadratic problem. Another problem closer to ours was mentioned in [16]. The authors assumed that the given routing paths are those defined as the shortest in terms of the number of hops, and the aim was to determine a set of weights which is compatible with these paths (i.e., with the uniqueness guarantee).

In the rest of this section, some necessary conditions for the existence of compatible weights are derived. The problem of determining compatible weights is then formulated as a linear mathematical program possibly involving integer variables. Finally, some particular cases of graphs are studied in more detail.

3.1. Necessary conditions. Let $G = (V, E)$ denote the graph associated with the network. Denote by K the set of demands for which the routing paths are given. In our case, each demand is simply defined as a pair of nodes (a, b). If $(a, b) \in K$, then $\phi(a, b)$ denotes the routing path which is required to be the unique shortest path between a and b. The set of all routing paths is denoted by Φ. We assume that the set E of G is exactly the set of edges that are contained in at least one of the required routing paths.

The first necessary condition uses the so-called concept of suboptimality.

Definition 3.2 (suboptimality). *A set of routing paths K satisfies the suboptimality condition if, for all pairs of demands (a, b) and (c, d) in K having two vertices e and f in common, then $\phi(a, b)$ and $\phi(c, d)$ share the same subpath between e and f.*

The suboptimality condition for two paths is illustrated in Figure 5.

Proposition 3.3 (first necessary condition). *If a set of weights is compatible with a given set of routing paths, then the set of paths satisfies the suboptimality condition.*

Proof. If the suboptimality condition is not satisfied for a pair of routing paths $\phi(a, b)$ and $\phi(c, d)$, then it is clear that they cannot both be unique shortest paths. □

In the rest of this section, we assume that the suboptimality condition is satisfied. Note that, if x and y are two vertices of $\phi(a,b)$, the shortest path between x and y must be the subpath of $\phi(a,b)$ linking x and y. Thus, we assume that $\phi(x,y)$ is also given and that there is a demand between x and y.

Unfortunately, the suboptimality condition is not sufficient to guarantee the existence of a compatible set of weights. A more elaborate condition can be derived using the notion of cyclic compatibility.

Definition 3.4 (cyclic compatibility). *A set of routing paths K is said to satisfy the cyclic compatibility condition if each edge e, which is not a bridge, is contained in at least one cycle C such that all demands $(a,b) \in K$ having both end-points in C are routed exclusively on edges of C, i.e., $\phi(a,b) \subset C$.*

In other words, the edge e belongs to at least one cycle which routes all the demands between its vertices. This condition is necessary for a set of weights to be compatible with a set of routing paths.

Theorem 3.5 (second necessary condition). *If a set of weights is compatible with a given set of routing paths, then the set of paths satisfies the cyclic compatibility condition.*

Proof. Let e be an edge which is not a bridge. e exists only if the graph is not a forest. Note that if the graph is a forest, then there is obviously a set of weights compatible with the routing paths.

Suppose that there is a set of weights which is compatible with the set of routing paths. Let C be a shortest cycle (in the sense of these weights) containing the edge e. Suppose that a and b are two vertices of C such that $(a,b) \in K$. Let $P_1(a,b)$ (resp., $P_2(a,b)$) denote the subpath of C between a and b containing (resp., not containing) e. First, assume that $\phi(a,b)$ contains e. If $\phi(a,b) \neq P_1(a,b)$, then by uniqueness of shortest paths the length of $P_1(a,b)$ is strictly higher than the length of $\phi(a,b)$. Thus, we can replace $P_1(a,b)$ by $\phi(a,b)$ to obtain a new cycle containing e whose length is lower than the length of C. As this is impossible by the definition of C, we can deduce that $\phi(a,b) = P_1(a,b)$ if $e \in \phi(a,b)$. In the same way, we can show that $\phi(a,b) = P_2(a,b)$ if $e \notin \phi(a,b)$. □

Figure 6 shows a graph and some routing paths satisfying the suboptimality condition. However, the cyclic compatibility is violated by edge 56. Indeed, any cycle C containing the edge 56 should contain only one of the two vertices 1 and 2. Suppose that C includes the vertex 1. If C satisfies the property of Definition 3.4, then it must include the vertex 4 because $4 \in \phi(1,5)$. As $2 \in \phi(6,4)$, then the vertex 2 is also in C. This is impossible because C cannot contain at the same time 1 and 2. By symmetry, we can show that C does not include the vertex 2. In other words, there is no cycle C containing 56 and satisfying the property of cyclic compatibility.

In fact, even if the two previous conditions are satisfied, the existence of a set of weights which is compatible with the routing paths is not guaranteed. The graph of Figure 7, obtained from the previous one by insertion of a vertex 7 between the vertices 5 and 6 and adding an edge 73, satisfies the two conditions. However, there is no set of weights satisfying the requirements. Another condition which is an extension of cyclic compatibility is violated here.

Definition 3.6 (generalized cyclic compatibility). *A set of routing paths K is said to satisfy the generalized cyclic compatibility condition if, for every pair of vertices (a,b) and every subset of edges $E' \subset E$ such that a and b are connected in $G \setminus E'$, there is at least one path p in $G \setminus E'$ between a and b satisfying the following property: if $c \in p$, $d \in p$, and $(c,d) \in K$, then either $\phi(c,d) \subset p$ or $\phi(c,d) \cap E' \neq \emptyset$.*

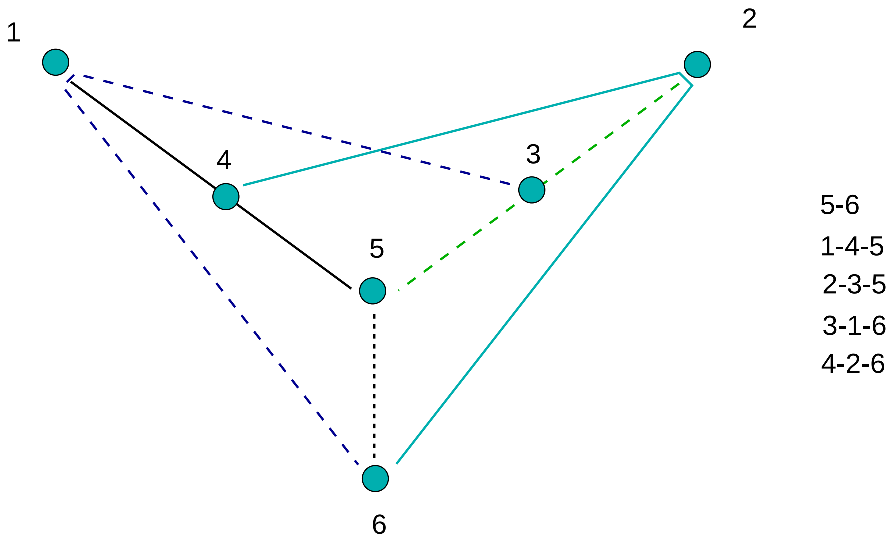


FIG. 6. *Cyclic compatibility violated by edge* 56.

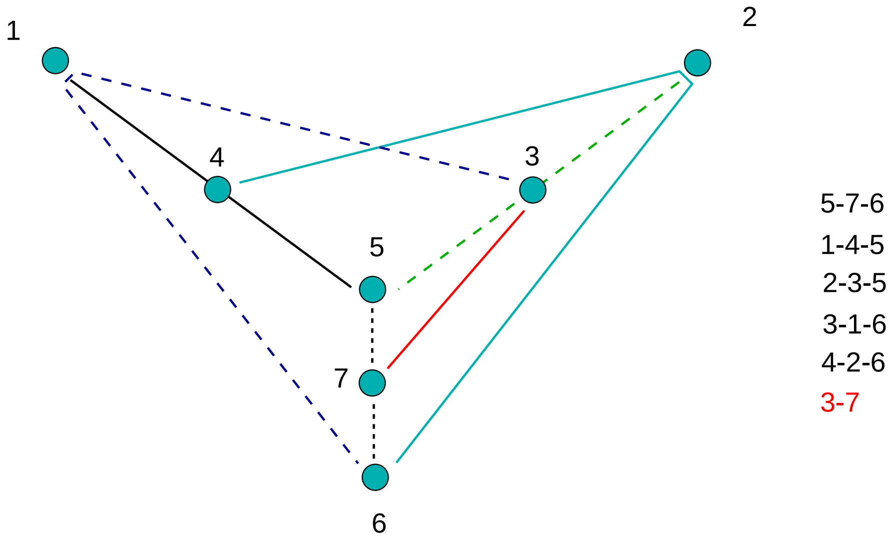


FIG. 7. *Generalized cyclic compatibility is violated ($E' = \{57, 76\}$ and $(a, b) = (5, 6)$).*

THEOREM 3.7 (third necessary condition). *If a set of weights is compatible with a given set of routing paths, then the set of paths satisfies the generalized cyclic compatibility condition.*

Proof. Let (a, b) be a pair of vertices and $E' \subset E$ a set of edges such that a and b are connected in $G \setminus E'$. Assume that there is a set of weights which is compatible with the set of routing paths. Let p be a shortest path in $G \setminus E'$ linking a and b. Let c and d be any two vertices of p such that $(c, d) \in K$. If $\phi(c, d) \cap E' = \emptyset$, then $\phi(c, d) \subset p$, otherwise p can be replaced by another path which is strictly shorter by uniqueness of shortest paths. Thus, we can deduce that $\phi(c, d) \subset p$ or $\phi(c, d) \cap E' \neq \emptyset$. □

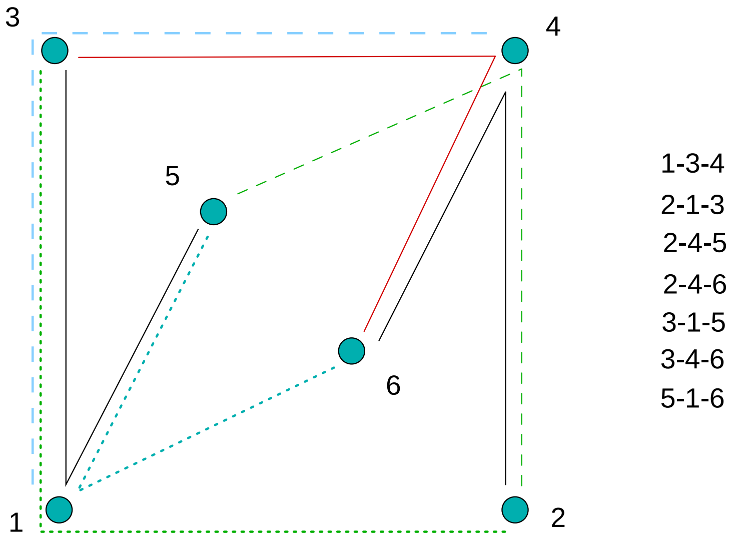


Fig. 8. *No possible compatible weights.*

It is easy to see that the generalized cyclic compatibility is violated by the routing paths of Figure 7. If we take $(a, b) = (5, 6)$ and $E' = \{57, 76\}$, then there is no path p between 5 and 6 satisfying the property given in Definition 3.6.

In fact, even if the three necessary conditions are satisfied, the existence of a set of weights is not guaranteed. The graph of Figure 8 satisfies the three conditions. However, there is no set of weights compatible with the routing paths. This is shown by solving the linear programs presented in the next section.

3.2. Linear programs to compute the weights. We define $S(a, b)$ as the set of elementary paths P between a and b other than $\phi(a, b)$. A variable y_{ab} gives the weight of $\phi(a, b)$. In the following two sections, we first focus on real weights, then we give methods to obtain integer weights. Throughout this paper, we assume that the weights are ≥ 1.

3.2.1. Real weights. The problem of computing a set of weights which is compatible with the routing paths ϕ can be formulated in a natural way as a linear program.

$$LP_1 = \begin{cases} \quad\text{Find} \quad (w_e)_{e\in E} \\ \text{subject to} \\ \sum_{e\in\phi(a,b)} w_e = y_{ab} \ \forall (a, b) \in K, \\ \quad \sum_{e\in P} w_e \geq 1 + y_{ab} \ \forall (a, b) \in K, \forall P \in S(a, b), \\ \quad\quad w_e \geq 1 \ \forall e \in E; \ \ y_{ab} \geq 0 \ \forall (a, b) \in K. \end{cases} \tag{3.1}$$

By definition, we have $\sum_{e\in\phi(a,b)} w_e = y_{ab}$. The inequalities $\sum_{e\in P} w_e \geq 1 + y_{ab}$ express the fact that the weight of any path P is higher than the weight of $\phi(a, b)$. As the number of constraints of LP_1 can be very high, we use a polynomial algorithm to generate the "useful" constraints. Thus, at each iteration, we add new constraints and we solve the new augmented linear program (see, for example, [29, 31] for a general description of this kind of algorithm). When all the constraints are satisfied (even those that were not generated), we obtain a set of weights that is compatible with the routing paths. In some cases, for example, when the necessary conditions of the previous section are not satisfied, LP_1 has no solution.

For a set of weights given by solving the linear program at any iteration, we check if there is a violated constraint for a pair of vertices $(a, b) \in K$. This is achieved by computing the two shortest elementary paths between a and b, in the sense of weights $(w_e)_{e \in E}$. We can use the Yen algorithm [21, 39], whose complexity is $O(n^3)$ or the algorithm of [25], whose complexity is lower than $\min\big[O(n^2), O(m \log(n))\big]$, where m is the number of edges and n the number of vertices. We can generate the constraints for all the pairs of vertices by applying one of the two algorithms to every pair of vertices. Another solution consists of using the algorithm of Dreyfus [14], which computes the two shortest paths (not necessarily elementary) between all the pair of vertices in fewer than $O(n^3)$ operations. Let us see exactly how can we generate a violated constraint by computing two shortest paths (not necessarily elementary) between two vertices a and b.

PROPOSITION 3.8. *For any set of weights, if there are violated constraints in the form of $\sum_{e \in P} w_e \geq 1 + y_{ab}$, then we can generate one of them by computing the two shortest paths (not necessarily elementary) between a and b.*

Proof. Let $T(a, b)$be the shortest path between a and b in the sense of the weights given by solving the linear program of the current iteration. As the weights are strictly positive, $T(a, b)$ is elementary. If $T(a, b) \neq \phi(a, b)$, then the constraint $\sum_{e \in T(a,b)} w_e \geq 1 + y_{ab}$ is violated and can be added to the current linear program. Next assume that $T(a, b) = \phi(a, b)$.

In this case, let $T'(a, b)$ be the second shortest path between a and b. If $T'(a, b)$ is elementary, then two cases have to be considered. If $\sum_{e \in T'(a,b)} w_e \geq 1 + y_{ab}$, then we are sure that all the constraints $\sum_{e \in P} w_e \geq 1 + y_{ab}$ are satisfied. Otherwise, the inequality $\sum_{e \in T'(a,b)} w_e \geq 1 + y_{ab}$ is violated and can be added to the current linear program.

Now, let us suppose that $T'(a, b)$ is not elementary, and let $T''(a, b)$ be the path obtained from $T'(a, b)$ by eliminating the cycles. If $T''(a, b) \neq \phi(a, b)$, then the weight of $T''(a, b)$ is strictly lower than the weight of $T'(a, b)$, which leads to a contradiction with the assumption that $T'(a, b)$ is a the second shortest path. Otherwise, $T''(a, b) = \phi(a, b)$, which means that $T'(a, b)$ contains at least two more edges in addition to those of $\phi(a, b)$. As the weights are higher than 1, the constraint $\sum_{e \in T'(a,b)} w_e \geq 1 + y_{ab}$ is satisfied. Consequently, all the constraints $\sum_{e \in P} w_e \geq 1 + y_{ab}$ are also satisfied. □

As the separation problem is polynomial, the linear program LP_1 can be solved in polynomial time using the ellipsoid algorithm [22]. Even if this remark has only a theoretical interest, the number of constraints that are added by constraint generation is generally not very large when the problem size is moderate (see the computational results section).

Since the problem LP_1 does not necessarily have a solution, it can be more interesting to solve the problem LP_2 defined below.

$$(3.2) \qquad LP_2 = \left\{ \begin{array}{l} \text{Maximize} \quad \sum_{(a,b) \in K} d_{ab} \times \epsilon_{ab} \\ \text{subject to} \\ \sum_{e \in \phi(a,b)} w_e = y_{ab} \ \forall (a, b) \in K, \\ \sum_{e \in P} w_e \geq \epsilon_{ab} + y_{ab}; \ \forall (a, b) \in K \ \forall P \in S(a, b), \\ 0 \leq \epsilon_{ab} \leq 1 \ \forall (a, b) \in K, \\ w_e \geq 0 \ \forall e \in E, \ y_{ab} \geq 0 \ \forall (a, b) \in K. \end{array} \right.$$

For every pair of vertices $(a, b) \in K$, a factor d_{ab} expresses the importance to satisfy the constraints regarding the routing of (a, b). A variable $0 \leq \epsilon_{ab} \leq 1$ indicates if the constraints are satisfied. Thus, $\sum_{(a,b)\in K} d_{ab} \times \epsilon_{ab}$ gives the amount of the total satisfaction of the constraints. Obviously, if LP_1 has a solution, then LP_2 also has a solution where all the variables ϵ_{ab} are equal to 1. Note that in LP_2 formulation we do not require the weights to be greater than 1. Some protocols can handle weights equal to 0. It is, however, possible to add to LP_2 the constraints $w_e \geq 1$ or any other linear constraints related to weights. A nice and important property of LP_2 (as formulated above), is that one can get an optimal solution by linear programming containing only 0/1 variables ϵ_{ab}. Said another way, a demand will be either completely satisfied ($\epsilon_{ab} = 1$) or not satisfied ($\epsilon_{ab} = 0$). The proof of this property is straightforward.

In [16], another linear formulation of the problem was given. This formulation contains $O(n^2)$ variables and $O(n^3)$ ($n = |V|$) constraints and is equivalent to LP_1. Next, we give a formulation using the same variables as those of [16] but slightly fewer constraints.

If the suboptimality condition is not satisfied, we know by Proposition 3.3 that there is no set of weights which is compatible with the routing paths. This condition can be checked in a polynomial time. Henceforth, we will assume that the suboptimality condition is satisfied.

A variable y_{ab} is associated with each pair of vertices of G, even if $(a, b) \notin K$.

We define $\overline{K} = \{(a, b) \in V \times V$ such that $\exists(c, d) \in K \ / \ \{a, b\} \subset \phi(c, d)\}$. Note that $\{a, b\} \subset \phi(c, d)$ means here that both vertices a and b are in the path $\phi(c, d)$. We have clearly $K \subset \overline{K}$. In other terms, $\overline{K}$ is the set of pairs for which the routing path is given in advance or can be directly deduced by suboptimality conditions from other given paths. In both cases, we will continue to use $\phi(a, b)$ to denote the routing path for $(a, b) \in \overline{K}$.

The new linear problem is given below.

$$(3.3) \quad LP_3 = \begin{cases} \qquad \text{Find} \quad (w_e)_{e\in E}, (y_{ab})_{(a,b)\in\overline{K}} \\ \text{subject to} \\ \sum_{e\in\phi(a,b)} w_e = y_{ab} \ \forall(a, b) \in \overline{K}, \\ \quad y_{ab} + 1 \leq w_{ac} + y_{cb} \ \forall(a, b) \in \overline{K}, \ \forall c \notin \phi(a, b), ac \in E, \\ \qquad y_{ab} \leq w_{ac} + y_{cb} \ \forall(a, b) \notin \overline{K}, \ \forall c \in V, ac \in E, \\ w_e \geq 1 \ \forall e \in E, \qquad y_{ab} \geq 0 \ \forall(a \neq b), \ y_{aa} = 0 \ \forall a \in V. \end{cases}$$

Recall that the demands are undirected. However, we consider the constraints $y_{ab}+1 \leq w_{ac}+y_{cb}$ in both directions. In other terms, the constraints $y_{ba}+1 \leq w_{bd}+y_{da}$ (where $bd \in E$ and $d \notin \phi(a, b)$) are also considered in LP_3.

One can easily see that the number of constraints of the linear program is $O(|E||V|)$.

Proposition 3.9. *The number of variables of LP_3 is $O(|V|^2)$ and the number of constraints is $O(|V||E|)$.*

Note that LP_1 can be seen as the projection of LP_3 onto a smaller subspace. The number of facets of LP_1 seems to be difficult to estimate, but the number of facets of LP_3 is polynomial. In fact, there are many examples of polyhedrons with exponentially many facets which can be represented as a projection of polyhedrons in higher dimensions but with a polynomial number of facets.

If the suboptimality condition is not satisfied, we know by Proposition 3.3 that there is no set of weights which is compatible with the routing paths. This condition can be checked in a polynomial time. Henceforth, we will assume that the suboptimality condition is satisfied.

THEOREM 3.10. *When the suboptimality condition is satisfied, LP_1 and LP_3 are equivalent.*

Proof. Consider a solution of LP_1 and let us build a solution of LP_3. If $(a,b) \in K$, then y_{ab} is directly given by the value of the LP_1 solution. Otherwise, y_{ab} is defined as the weight of the shortest path between a and b in the sense of the weights of (LP_1). Recall that $G = (V,E)$ is supposed to be connected so y_{ab} is well defined. Let us see if the constraints of (LP_3) are satisfied. If $(a,b) \in \overline{K}$, $ac \in E$, and $c \notin \phi(a,b)$, then $w_{ac} + y_{cb}$ is the weight of a path between a and b other than $\phi(a,b)$. As the weights are a solution of (LP_1), $w_{ac} + y_{cb}$ is necessarily higher than $1 + y_{ab}$. When $(a,b) \notin \overline{K}$, the inequalities $y_{ab} \leq w_{ac} + y_{cb}$ are satisfied by definition of the variables y. Thus, we built a solution of (LP_3) using a solution of (LP_1).

Now, let us consider a solution of LP_3. We will show that it induces a solution of LP_1. In other words, we want to prove that $\sum_{e\in P} w_e \geq 1 + y_{ab}$ for any pair $(a,b) \in K$ and any path $P \in S(a,b)$. This property will be shown by induction. More precisely, we consider the following property: for any path P containing at most i edges and linking a and b, if $(a,b) \in \overline{K}$ and $P \neq \phi(a,b)$, then $\sum_{e\in P} w_e \geq 1 + y_{ab} = 1 + \sum_{e\in\phi(a,b)} w_e$, otherwise $\sum_{e\in P} w_e \geq y_{ab}$.

Let us begin with $i = 1$. If $(a,b) \in \overline{K}$ and the length of P is 1, then $P = ab = \phi(a,b)$ (by suboptimality and the definition of E). Thus, there is not any path $P \neq \phi(a,b)$ of length 1. If $(a,b) \notin \overline{K}$ and the length of P is 1, then $P = ab$ and the inequality $w_{ab} \geq y_{ab}$ is satisfied because it is one of the constraints of LP_3.

Now, suppose that the property is valid for $i-1$ and let us show it for i. Let P be a path containing exactly i edges, if there is any. First, focus on the pairs (a,b) of $\overline{K}$. If $P \neq \phi(a,b)$ contains only vertices from $\phi(a,b)$, then we can deduce by the suboptimality condition that P contains in fact all the edges of $\phi(a,b)$ and is cyclic. As the weights are higher than 1, we have immediately $\sum_{e\in P} w_e \geq 1 + y_{ab}$. On the other hand, if P contains at least one vertex that is not in $\phi(a,b)$, then we define c as the first vertex not in $\phi(a,b)$ encountered when we go through p from a to b. Let d be the predecessor of c in P. Obviously, we have $d \in \phi(a,b)$. We can write that $P = P_1 \cup dc \cup P_2$, where P_1 is the subpath going from a to d and P_2 is a path going from c to b. Both P_1 and P_2 contain fewer than $i-1$ edges. By induction, in all the cases we have $\sum_{e\in P_1} w_e \geq y_{ad}$ and $\sum_{e\in P_2} w_e \geq y_{cb}$. Adding up these two inequalities leads to $\sum_{e\in P} w_e \geq y_{ad} + w_{dc} + y_{cb}$. Since the weights are a solution of LP_3, $w_{dc} + y_{cb} \geq y_{db} + 1$. Consequently, $\sum_{e\in P} w_e \geq y_{ad} + y_{db} + 1$. Moreover, we know that $d \in \phi(a,b)$. This implies that both (a,d) and (d,b) are in $\overline{K}$. In other words, $y_{ad} + y_{db} = y_{ab}$. Combining this equality with the previous inequality leads to $\sum_{e\in P} w_e \geq y_{ab} + 1$.

Finally, suppose that $(a,b) \notin \overline{K}$ and let c be the last vertex of P encountered before b. If $c = a$, then a and b are adjacent and P contains this edge. This means that $\sum_{e\in P} w_e \geq w_{ab} \geq y_{ab}$. Otherwise, let P' be the subpath of P linking a and c. This path has exactly $i-1$ edges, so by the induction hypothesis we can deduce that $\sum_{e\in P'} w_e \geq y_{ac}$. This implies that $\sum_{e\in P} w_e \geq y_{ac} + w_{cb}$. On the other hand, the weights are a solution of LP_3, which leads to$y_{ba} \leq w_{bc} + y_{ca}$. Consequently, the inequality $\sum_{e\in P} w_e \geq y_{ab}$ is again valid.

Thus, the property is proved and every solution of LP_3 induces a solution of LP_1. □

Note that even if the number of constraints and variables of LP_3 is polynomial, solving LP_3 can be more time consuming than solving LP_1.

Remarks. Some of the linear programs presented in this section (LP_1 and LP_2) can easily be extended to integrate many other technical constraints.

- Traffic demands can be routed using many paths to improve the use of network capacities. Thus, we can assume that n paths are given (instead of one path) for every demand and must be the n shortest paths between the extremities of the traffic demand. This requirement can be expressed very easily by adding linear constraints (such as those of LP_1), which are determined by solving some n-shortest path problems.
- Sometimes, due to transmission delay limitations and the cost of the routing function, the paths having more than a given number of hops are eliminated. Thus, the constraints of LP_1 ($\sum_{e\in P} w_e \geq 1 + y_{ab}$) are needed only for paths having a limited number of hops. The violated constraints can again be determined by solving the 2-shortest path problem, but we need to slightly modify Yen's or Dreyfus's algorithms to integrate the hop constraint.
- The routing paths can be the output of a dimensioning procedure [4]. If we consider survivability, we can also have imposed paths when there is an edge or a node failure. The same kind of linear constraints are added to satisfy this requirement. We only have to delete the failing edge (or the edges incident to a failing node) and to apply the same algorithms on the new graph.
- We assumed that E is the set of edges belonging to at least one of the given routing paths. In fact, it is possible to consider some other edges. This can be done in LP_1 by adding variables w_e corresponding to edges that are not in E. We can also solve LP_3 and fix the weight w_e of an edge $e = ab \notin E$ as $w_e = y_{ab} + 1$.
- The set of weights that are compatible with the routing path of a demand is a nonempty polyhedron. Moreover, a polyhedron is a convex set.
  Let us recall the well-known theorem of Helly. This theorem states that if Ω is a finite family of convex sets in $\mathbb{R}^d$ such that any $d+1$ or fewer of the sets of Ω intersect, then $\cap\Omega \neq \emptyset$. This theorem can be applied here where $d = |E|$, Ω is the set of polyhedrons associated with each demand. By Helly's theorem, we can deduce that we can find a set of weights compatible with all the traffic demands if and only if there exists a solution for any $|E|+1$ or fewer traffic demands. In other words, the compatibility of all sets containing $|E|+1$ or fewer demands implies the existence of a set of weights which is compatible with all demands.
  A stronger result can be obtained by considering another family of convex sets. For any demand $(a,b) \in K$ and any path $P \in S(a,b)$, we define $\gamma(a,b,P)$ as the set of weights such that $\sum_{e\in P} w_e \geq 1 + \sum_{e\in\phi(a,b)} w_e$ and $w_e \geq 1$ for all edges. The sets $\gamma(a,b,P)$ are convex sets. Helly's theorem implies that the compatibility problem has a solution if and only if any $|E|+1$ or fewer of the sets $\gamma(a,b,P)$ intersect. It is also possible to add lower and upper bounds for the weights without breaking the convexity of the sets $\gamma(a,b,P)$. Then, the property described above is still valid.

3.2.2. Integer weights. An important constraint that is imposed by some protocols is the integrality of weights. In this section, we will focus on methods that can be used to find small integer weights. The set of integers is denoted $\mathbb{N}$. An obvious way to obtain this kind of weights consists in solving the following integer program IP_4.

$$(3.4)\qquad IP_4 = \begin{cases} \text{Minimize} \quad w_{max} \\ \text{subject to} \\ \sum_{e\in\phi(a,b)} w_e = y_{ab}\ \forall(a,b)\in K, \\ \quad \sum_{e\in P} w_e \geq 1 + y_{ab}\ \forall(a,b)\in K, \forall P\in S(a,b), \\ \quad 1 \leq w_e \leq w_{max}\ \forall e\in E, \\ \qquad w_e\in\mathbb{N}\ \forall e\in E,\ y_{ab}\geq 0\ \forall(a,b)\in K. \end{cases}$$

It is also possible to provide another integer model by adding the integrality constraint to LP_3 instead of LP_1. Let I be the value of the objective function of IP_4. Let LP_4 be the continuous relaxation of IP_4. We also define C as the value of the objective function when LP_4 is solved.

We obviously have $\lceil C\rceil \leq I$. Thus, a lower bound for I is easily obtained by solving LP_4. Moreover, an upper bound can also be deduced from this relaxation. Indeed, if B is the basis matrix used at an optimum when LP_4 is considered, then it is clear that we can multiply the weights by $\det(B)$ (determinant of B) to obtain integer weights. These weights can be reduced by dividing them by the greatest common divider which can be computed using the Euclidean algorithm. This method is easy and gives an upper bound of I. However, the determinant can be high and the quality of the bound is not always good.

Below we give another simple algorithm to compute integer weights. We will use again the set $\overline{K}$ defined in the previous section as the set of pairs for which the routing path is given in advance or can be directly deduced by suboptimality conditions from other given paths. Let $D(a,b)$ be the set of paths between a and b which are edge disjoint with $\phi(a,b)$. The set of real numbers is denoted $\mathbb{R}$.

A new linear program LP_5 is defined below.

$$(3.5)\quad LP_5 = \begin{cases} \text{Minimize} \quad w_{max} \\ \text{subject to} \\ \sum_{e\in\phi(a,b)} w_e = y_{ab}\ \forall(a,b)\in \overline{K}, \\ \quad \sum_{e\in P} w_e \geq y_{ab} + \frac{|P|+|\phi(a,b)|}{2}\ \forall(a,b)\in \overline{K}, \forall P\in D(a,b), \\ \quad 0.51 \leq w_e \leq w_{max}\ \forall e\in E, \\ \qquad w_e\in\mathbb{R}\ \forall e\in E\ y_{ab}\geq 0,\ \forall(a,b)\in K. \end{cases}$$

We used in LP_5 the notation $|P|$ to designate the number of edges of a path P. To obtain integer weights, we propose the following simple heuristic.

Upper bound heuristic.

1. Solve LP_5.
2. Every edge weight is rounded to the nearest integer. If there is an ambiguity, it is rounded to the lowest integer.

Before showing how LP_5 can be solved, we have to prove the correctness of this algorithm.

PROPOSITION 3.11. *The solution given by the upper bound heuristic is a feasible solution of IP_4.*

Proof. Let $(w_e)_{e\in E}$ be an optimal solution of LP_5 and $(w'_e)_{e\in E}$ the final solution obtained after the rounding procedure.

Let us consider a pair of vertices $(a,b)\in\overline{K}$ and a path $P\in D(a,b)$. By LP_5 constraints, we have $\sum_{e\in P} w_e - \sum_{e\in\phi(a,b)} w_e \geq \frac{|P|+|\phi(a,b)|}{2}$. As $|w_e - w'_e| \leq 0.5$ for every edge $e\in E$, we can deduce that $\sum_{e\in P} w'_e - \sum_{e\in\phi(a,b)} w'_e \geq 0$. Moreover, if $\sum_{e\in P} w'_e - \sum_{e\in\phi(a,b)} w'_e = 0$, then we necessarily have $w'_e = w_e + 0.5$ when $e\in\phi(a,b)$ and $w'_e = w_e - 0.5$ when $e\in P$. But we have decided to round the weights to the lower integer when there is an ambiguity. Thus we cannot have $\sum_{e\in P} w'_e - \sum_{e\in\phi(a,b)} w'_e = 0$, which implies that $\sum_{e\in P} w'_e - \sum_{e\in\phi(a,b)} w'_e \geq 1$. This is valid for any $(a,b) \in \overline{K}$ and any path $P \in D(a,b)$. We can easily see that this leads to $\sum_{e\in P} w'_e - \sum_{e\in\phi(a,b)} w'_e \geq 1$ for any $(a,b) \in K$ and any $P \in S(a,b)$.

The weights $(w'_e)_{e\in E}$ are clearly higher than 1. All the constraints of IP_4 are satisfied and the correctness is proved. □

Let U be the value of the highest weight given by the upper bound heuristic. We clearly have $\lceil C\rceil \leq I \leq U$. The next proposition gives a guarantee on the worst case performance of our heuristic. We define ϕ_{max} as follows: $\phi_{max} = \max_{(a,b)\in\overline{K}} |\phi(a,b)| = \max_{(a,b)\in K} |\phi(a,b)|$. We also use N to denote the number of nodes.

THEOREM 3.12. $U \leq \lceil C\rceil \min(\frac{N}{2}, \phi_{max}) \leq I.\min(\frac{N}{2}, \phi_{max})$.

Proof. Let U_5 be the maximum weight of the optimal solution of LP_5. Therefore, we have $U < U_5 + 0.5$. Moreover, using the fact that P and $\phi(a,b)$ are edge disjoint, we can deduce that $\frac{|P|+|\phi(a,b)|}{2} \leq \frac{N}{2}$. This implies that any solution of LP_4 (continuous relaxation of IP_4) multiplied by $\frac{N}{2}$ becomes a feasible solution of LP_5. Consequently, we necessarily have $U_5 \leq \frac{N}{2}C$ which leads to $U < \frac{N}{2}C + 0.5 \leq \frac{N}{2}\lceil C\rceil + 0.5$. As U is an integer, we can deduce that $U \leq \frac{N}{2}\lceil C\rceil$.

Let LP_6 be the linear program presented below:

$$LP_6 = \begin{cases} \text{Minimize} \quad w_{max} \\ \text{subject to} \\ \sum_{e\in\phi(a,b)} w_e = y_{ab} \ \forall (a,b) \in \overline{K}, \\ \quad \sum_{e\in P} w_e \geq y_{ab} + |\phi(a,b)| \ \forall (a,b) \in \overline{K} \ \forall P \in D(a,b), \\ \quad 0.01 \leq w_e \leq w_{max} \ \forall e \in E, \\ \qquad w_e \in \mathbb{R} \ \forall e \in E, \ y_{ab} \geq 0 \ \forall (a,b) \in K. \end{cases} \tag{3.6}$$

This program can be obtained from LP_5 by changing w_e to $w_e + 0.5$. Let U_6 be the maximum weight of the optimal solution of LP_6. We clearly have $U_5 = U_6 + 0.5$. Moreover, any solution of LP_4 multiplied by ϕ_{max} becomes a feasible solution of LP_6. Consequently, $U_6 \leq \phi_{max}C$. This leads to $U_5 \leq \phi_{max}\lceil C\rceil + 0.5$. Combining the previous inequality with the inequality $U < U_5 + 0.5$ implies that $U < \phi_{max}\lceil C\rceil + 1$. As U is an integer, we can deduce that $U \leq \phi_{max}\lceil C\rceil$. □

Note that the worst-case guarantee of the previous theorem can be reached in some cases. This occurs, for example, when the graph is a Hamilton cycle with an odd number of vertices and the routing paths are the shortest paths in sense of number of hops. Indeed, we have $I = 1$ and $U = \lfloor\frac{N}{2}\rfloor = I.\min(\frac{N}{2}, \phi_{max})$.

However, it is sometimes possible to improve the results of the upper bound heuristic. To do this, we only have to decrease by 1 all the highest weights until some constraints of IP_4 become violated. In the case of the Hamilton cycle described above, we obtain weights equal to 1 if we proceed in this way.

Now, we have to show how LP_5 can be solved. In fact, it seems to be easier to first solve LP_6 and then to change the weights w_e to $w_e + 0.5$. LP_6 can clearly be solved by constraint generation. We first compute the shortest path P between a and b for every pair of vertices $(a,b) \in \overline{K}$. If the shortest path P is different from $\phi(a,b)$, then we add the constraint $\sum_{e\in P} w_e \geq y_{ab} + |\phi(a,b)|$. If not, we compute the shortest path P' in the graph obtained by eliminating the edges of $\phi(a,b)$. The constraint $\sum_{e\in P'} w_e \geq y_{ab} + |\phi(a,b)|$ is added if it is violated.

In the next section, we will give some important particular graphs for which the suboptimality condition is sufficient to find a set of weights compatible with the routing paths.

3.3. Particular graphs. First, recall that the set of edges E of the graph G is exactly the set of edges that are contained in at least one routing path of a demand of K. We assume that the suboptimality condition (Proposition 3.3) is satisfied, and we focus on particular kinds of graphs.

LEMMA 3.13. *If $G = (V, E)$ is a cycle and the suboptimality condition is satisfied, then for any pair of vertices (a, b) which is not in K, it is possible to find a path between a and b not violating the suboptimality condition.*

Proof. As the graph is a cycle, there are only two simple paths between a and b. If one of them violates the suboptimality condition, then it necessarily contains a pair of nodes (c, d) of K such that $a \in \phi(c, d)$ and $b \in \phi(c, d)$. Consequently, the other path between a and b is a subpath of $\phi(c, d)$. As the suboptimality condition is satisfied by the routing paths of K, we can choose this path between a and b without violating the condition. □

Using the previous lemma, we can assume that K includes all the pair of vertices of G. Even if this is not necessary, we can define a routing path between every pair of vertices. Any set of weights that is compatible with all the routing paths is also compatible with any subset of these routing paths.

PROPOSITION 3.14. *Let $G = (V, E)$ be a cycle. The routing paths are assumed to satisfy the suboptimality condition. Suppose that for any edge ab, there exists at least one vertex x such that $\phi(x, a) \cap \phi(x, b) = \emptyset$. Then, the number of vertices n is odd and all the demands are routed on the minimum hop paths.*

Proof. Let y be any vertex of G. The cycle is clockwise oriented.

Let y' (resp., y") be the first vertex that precedes (resp., follows) y. Let z be the first vertex encountered when going through the cycle (starting from y) such that $y' \notin \phi(y, z)$. This vertex exists because $y' \notin \phi(y, y'') = yy''$. We also define z'' as the vertex that precedes. Clearly, we have $\phi(y, z) \cap \phi(y, z'') = \emptyset$. In other words, for any vertex y, there exists an edge zz" such that $\phi(y, z) \cap \phi(y, z'') = \emptyset$. This edge is unique by virtue of the suboptimality condition.

On the other hand, we know by the proposition hypothesis that for any edge ab, there is at least one vertex x such that $\phi(x, a) \cap \phi(x, b) = \emptyset$. Using the fact that the number of edges and the number of vertices are equal and combining the previous remarks leads to the fact that for any edge ab, there exists exactly one vertex x such that $\phi(x, a) \cap \phi(x, b) = \emptyset$. This vertex is denoted x_{ab}.

Without loss of generality, assume that a, b, and x_{ab} are located on the cycle as shown on Figure 9 (clockwise). Let c (resp., y) be the vertex which follows b (resp., x_{ab}). By definition of x_{ab}, we have $c \in \phi(x_{ab}, b)$ and $c \notin \phi(y, b)$. As we know that there exists a unique vertex x_{bc} such that $\phi(x_{bc}, b) \cap \phi(x_{bc}, c) = \emptyset$, we can deduce that $x_{bc} = y$. Consequently, the relative locations of a, b, and x_{ab} do not change by rotation.

Let d be the number of edges between x_{ab} and a (see Figure 9). Using the previous remark, this number does not depend on the identity of edge ab. Thus, the number of edges between b and x_{ab} is given by $n - d - 1$. If $d \geq n - 1 - d + 1$, then there exists an edge ab and an edge ef such that e, f, and x_{ef} are in $\phi(x_{ab}, a)$. But this is clearly impossible according to the suboptimality assumption. Thus, we necessarily have $2d \leq n - 1$. Similarly we can show that it is impossible to have $n - 1 - d \geq d + 1$, which means that $2d \geq n - 1$. In other words, n is odd and $d = \frac{n-1}{2}$. Using the

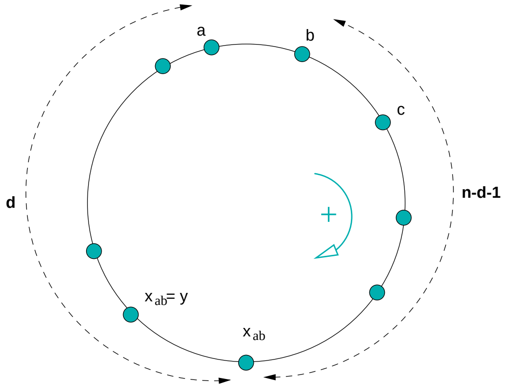


FIG. 9. *Cycle of Proposition* 3.14.

definition of x_{ab} and d, we can deduce that all the demands are routed on the shortest path in sense of the number of edges. □

Before giving the main theorem of this section, we need to provide some definitions. If $G = (V, E)$ is a cycle, we define $\delta_0(G, \phi)$ as the number of edges ab such that $\phi(x, a) \cap \phi(x, b) \neq \emptyset$ for any vertex $x \neq a, b$. For any set of weights $(w_e)_{e \in E}$ and any pair of vertices (x, y), $w\left(\phi(x, y)\right)$ denotes the weight of $\phi(x, y)$ and $w(\bar{\phi}(x, y))$ the weight of the complementary path. Recall that K is the set of all pairs of vertices of V.

THEOREM 3.15. *Let $G = (V, E)$ be a cycle. If the routing paths satisfy the suboptimality condition, then it is possible to determine a set of weights which is compatible with these routing paths. Moreover, an example of such a set of weights is given as follows: $w_{ab} = 1$ if $\phi(x, a) \cap \phi(x, b) \neq \emptyset$ for any vertex $x \neq a, b$ and $w_{ab} = w_0$ else, where w_0 is a constant number such that $w_0 \geq \delta_0(G, \phi) + 1$. Using these weights, we have the following property:* $\min_{(x,y) \in K} \left(w\left(\bar{\phi}(x, y)\right) - w\left(\phi(x, y)\right)\right) \geq w_0 - \delta_0(G, \phi)$.

Proof. We will show by induction on $\delta_0(G, \phi)$ that the weights given as described in the theorem satisfy the property $\min_{(x,y) \in K} \left(w\left(\bar{\phi}(x, y)\right) - w\left(\phi(x, y)\right)\right)$ $\geq w_0 - \delta_0(G, \phi) \geq 1$. First, if $\delta_0(G, \phi) = 0$, then we know by Proposition 3.14 that the demands are routed through the minimum hops path. This means that $\min_{(x,y) \in K} \left(w\left(\bar{\phi}(x, y)\right) - w\left(\phi(x, y)\right)\right) \geq w_0$. Suppose that the property is satisfied when $\delta_0(G, \phi) \leq l_0 - 1$, and let us show it is also satisfied for $\delta_0(G, \phi) = l_0 > 1$. Let ab be an edge of G such that $\phi(x, a) \cap \phi(x, b) \neq \emptyset$ for any vertex $x \neq a, b$. This is equivalent to saying that $\phi(x, a) \subset \phi(x, b)$ or $\phi(x, b) \subset \phi(x, a)$. Let a' and b' be the other neighbors of a and b (Figure 10). Let $G' = (V', E')$ be the cycle obtained by contracting the edge ab into a vertex c. We will build a new set of routing paths ϕ' in G' based on the initial set of routing paths ϕ. For any vertices $x, y \neq c$ of V', $\phi'(x, y)$ is exactly the same as $\phi(x, y)$ but taking into account the contraction of edge ab. Let $x \neq c$ be any vertex of G'. We take $\phi'(x, c) = \phi(x, a') \cup a'c$ if $\phi(x, a) \subset \phi(x, b)$ and $\phi'(x, c) = \phi(x, b') \cup b'c$ otherwise. The suboptimality condition is satisfied by the new set of routing paths ϕ'. We obviously have $\delta_0(G', \phi') = \delta_0(G, \phi) - 1$. Let w_0 be any constant number such that $w_0 \geq \delta_0(G, \phi) + 1$, and let us consider a set of weights of G exactly as defined in the theorem. In the new graph G', we take the same weights as those of G for the edges xy when $x, y \neq c$. Otherwise, $w_{a'c} = w_{a'a}$ and $w_{b'c} = w_{b'b}$. As $w_0 \geq \delta_0(G, \phi) + 1 \geq l'(G', \phi') + 1$, we can deduce by induction hypothesis that

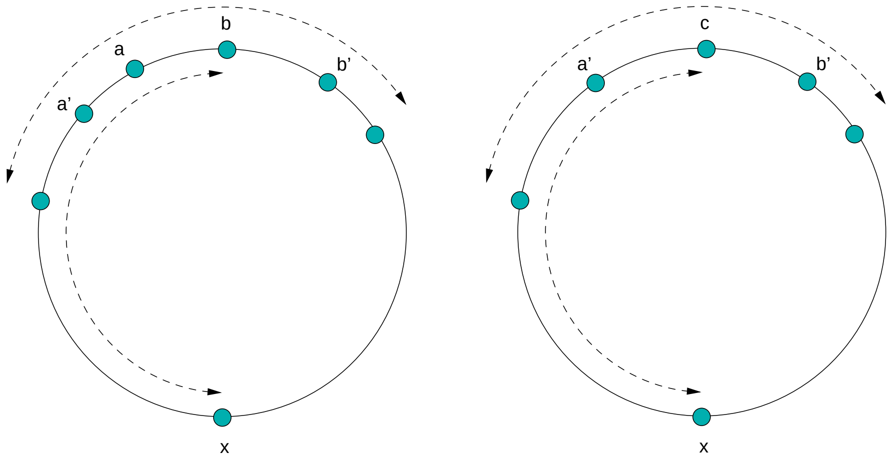


FIG. 10. *Contraction of edge ab.*

$\min_{(x,y)\in V'\times V'}\left(w\left(\bar{\phi}'(x,y)\right)-w\left(\phi'(x,y)\right)\right)\geq w_0-l'(G',\phi')$. As the weight of edge ab is 1, the previous inequality leads to $\min_{(x,y)\in V\times V}\left(w\left(\bar{\phi}(x,y)\right)-w\left(\phi(x,y)\right)\right)\geq w_0-l'(G',\phi')-1=w_0-\delta_0(G,\phi)$. □

To prove the above theorem, we contracted an edge ab for which $\phi(x,a)\cap\phi(x,b)\neq\emptyset$ for any vertex $x\neq a,b$. The number of edges ef for which there exists x such that $\phi(x,e)\cap\phi(x,f)=\emptyset$ does not change by contracting ab. Thus, if we contract all such edges we will obtain a cycle similar to those of Proposition 3.14. We can deduce that the number of edges ef for which there exists x such that $\phi(x,e)\cap\phi(x,f)=\emptyset$ is an odd number. Recall that all the cycle edges belong to at least one routing path. As any cycle having three vertices cannot contain any edge ab for which $\phi(x,a)\cap\phi(x,b)\neq\emptyset$ for every vertex $x\neq a,b$, we can deduce that the number of edges ef for which there exists x such that $\phi(x,e)\cap\phi(x,f)=\emptyset$ is at least equal to 3. These remarks are brought together in the following corollary.

Recall that we assumed that a routing path is given for every pair of vertices of the cycle.

COROLLARY 3.16. *Let $G=(V,E)$ be a cycle. If a routing path $\phi(a,b)$ is given for any pair of vertices of the cycle and if the suboptimality condition is satisfied, then the number of edges ef for which there exists x such that $\phi(x,e)\cap\phi(x,f)=\emptyset$, given by $n-\delta_0(G,\phi)$, is an odd number at least equal to 3.*

The suboptimality of the routing paths is sufficient to find a set of compatible weights for many other particular graphs. Some of them are defined below.

DEFINITION 3.17. *We define a hat-cycle as a graph which consists of an elementary cycle C and some vertices such that*

- every vertex x which is not in the cycle has exactly two neighbors; moreover, these two neighbors constitute an edge of C;

- for any edge ab of C, there is at most one vertex $x\notin C$ which is adjacent to both a and b.

An example of a hat-cycle is shown in Figure 11.

PROPOSITION 3.18. *Let $G=(V,E)$ be a hat-cycle. If the routing paths satisfy the suboptimality condition, then there exists a set of weights which is compatible with these routing paths.*

Proof. We will show the proposition by induction on the number of vertices that

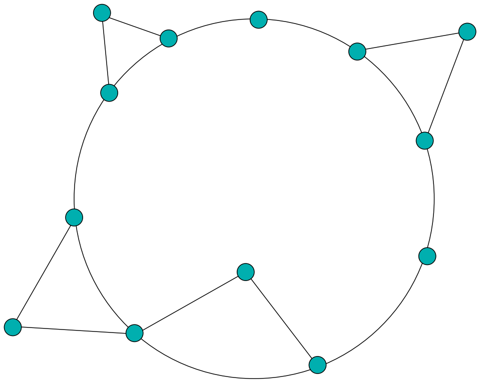


FIG. 11. *A hat-cycle.*

are not in C. If this number is nought, then the graph is a cycle and the result is given by Theorem 3.15. Suppose that the proposition is proved if this number is lower than $n_0 - 1$, and let us show it when this number is equal to n_0. Note that by definition of a hat-cycle, n_0 cannot be higher than the number of vertices of C. Let x be one of the vertices of $G\backslash C$ and y, z its two neighbors. By suboptimality, the routing path $\phi(y, z)$ is made up of the edge yz. Let G' be the graph obtained by eliminating the edge yz. We also define a new set of routing paths ϕ' in G' using the initial routing paths and replacing yz by $yx \cup xz$. Suboptimality is still satisfied. By induction hypothesis, we can find a set of weights $(w_e)\ _{e\in E}$ which is compatible with the routing paths ϕ'. Now, if we go back to our initial graph G, we can take for the edge yz a weight which is slightly lower than $w_{yx} + w_{xz}$. For the other edges, we choose the weights given by induction. All the constraints are satisfied and the proposition is shown. □

PROPOSITION 3.19. *If $G = (V, E)$ is a tree or a clique, then the suboptimality condition is sufficient to guarantee the existence of a set of compatible weights.*

Proof. In both cases, we can take a weight equal to 1 for all the graph links. □

This property is still satisfied even by some incomplete cliques as shown below.

PROPOSITION 3.20. *If $G = (V, E)$ is a graph of n vertices where the degree of any vertex is at least $n - 2$, then the suboptimality condition is sufficient to guarantee the existence of a set of compatible weights.*

Proof. Let (a, b) be any pair of vertices of G. If a and b are adjacent, then by suboptimality $\phi(a, b) = ab$. In the other case, both a and b are adjacent to all the other vertices of the graph. This is due to the fact that all the degrees are higher than $n - 2$. Using this remark and the suboptimality condition implies that $\phi(a, b)$ is made up of exactly two links ax and xb, where x is a vertex of G. In other words, all the routing paths of ϕ include either one link or two links. Let us define a set of weights as follows: if an edge xy does not belong to any two-link routing path, then $w_{xy} = 1$, else $w_{xy} = \frac{2}{3}$. If a and b are adjacent, then it is clear that the shortest path between them in the sense of these weights is unique and is given by the edge ab. Next assume that a and b are not adjacent. As said before, $\phi(a, b) = ax \cup xb$. The weights of both ax and xb are equal to $\frac{2}{3}$. Note that the weight of $\phi(a, b)$ is given by $\frac{4}{3}$ and there is clearly no other path between a and b which has a strictly lower weight. Thus, we have to show that there are no paths other than $\phi(a, b)$ having a weight equal to $\frac{4}{3}$. Let us prove this property by contradiction. Any other path between a and b whose weight is $\frac{4}{3}$ is necessarily made up two links ay and yb where

$y \neq x$. This implies that both ay and by are contained in some two-link routing paths of ϕ. Moreover, any two-link routing path containing ay has necessarily the form of some $\phi(a, z)$ or $\phi(z, y)$. As a and b are not adjacent, a is adjacent to all the other vertices. Thus, ay cannot be contained in any $\phi(a, z)$. Let us assume that $\phi(z, y)$ includes ay. This clearly means that z and y are not adjacent and y is adjacent to all the other vertices. We also know that b is adjacent to all the vertices other than a and $\phi(a, b)$ does not include by. Combining these two remarks implies that there is not any two-link routing path containing by. This contradicts the assumption that $w_{by} = \frac{2}{3}$. Thus, the shortest path between any two vertices a and b is always unique and is nothing other than the routing path $\phi(a, b)$. This proves the proposition and gives, by the way, an example of compatible weights. □

In fact, even if the graph is, in a certain sense, made up of some of the particular graphs described in this section, then it still satisfies the same property. This is established in the next theorem.

Recall that E is the set of edges used by at least one routing path $\phi(a, b)$ where $(a, b) \in K$. We will use in the next theorem the notion of a "block," which is defined as a bridge or a maximal (in sense of inclusion) two-connected induced subgraph of G [6, 12].

THEOREM 3.21. *If the suboptimality condition is satisfied, then it is possible to determine a set of compatible weights if every block of G belongs to one of the following graph classes: bridges, cycles, hat-cycles, cliques, and incomplete cliques (those of Proposition* 3.20*).*

Proof. This theorem is, in fact, a direct consequence of the propositions of this section and the definition of blocks. If G is not connected, then the result can be deduced by applying the theorem to all the connected components of G. If the graph is connected, then its block graph is a tree. We also know by definition that any two blocks have at most one common vertex [6, 12]. Using these remarks and what was shown in this section gives us the proof of the theorem. □

A sample graph similar to those described in the previous theorem is presented in Figure 12. Finally, note that the graphs called cactus [6] and defined as the connected graphs all of whose blocks are elementary cycles or bridges satisfy the condition of the previous theorem.

4. Computational results. The construction scheme proposed in section 2 was coded in C and run on a Sun Enterprize 450 with four 250 MAZE CPUs and 1 gigabyte of RAM. Some of the linear programs of section 3 were solved using a CPLEX Linear Optimizer 6.0 [10]. Random graphs with n vertices were obtained by generating either a given number m of edges or a given number of edges corresponding to a given graph density d. Note that density is often defined as the ratio $\frac{m}{n(n-1)/2}$. However, as we focus here on connected graphs, we use the following definition:

$$m = n - 1 + d * (n(n-1)/2 - (n-1)) .$$

Hence, a density of 0% corresponds to a tree, and a graph with a density 100% is a complete graph. In all cases, we make sure that the graph generated is connected (either by first generating a tree for the low density graphs or by using repeatedly the LEDA graph generator [28] until the connectivity requirement is satisfied).

The common objective of all the algorithms proposed in this paper is to obtain a set of weights such that all demands are routed on a single shortest path. In a first step, we used the construction scheme proposed in section 2 to build a breadth search tree and deduce the corresponding weights. As a byproduct, we also obtain a routing

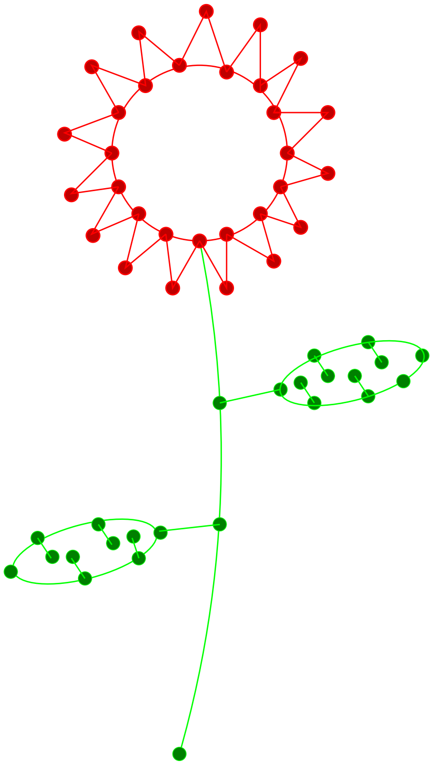


FIG. 12. *Suboptimality is sufficient to find compatible weights here.*

path (defined as the single shortest path) between each pair of nodes. One of the aims of the construction scheme is to obtain low value weights. In order to evaluate how well this aim is reached, we then use some of the linear programs of section 3 in which the objective is precisely, given a set of routing paths, to minimize the value of the maximum weight. Note that, in practice, it would be necessary to check if the solution set is not empty, i.e., if there exists a set of weights compatible with the routing paths. In the case of our experimental study, we already have a compatible set of weights. In our computational experiment, three mathematical programs are solved for each instance:

IP_4 is solved to obtain the reference optimal value I;
LP_4 is the continuous relaxation of IP_4 that gives a lower bound $\lceil C \rceil$;
LP_6 is solved to give the upper bound U (section 3.2.2).

Note that each run is limited to one hour of computing time. 10 instances are generated for each graph size n and each density d. When some instances could not be solved in this limited time, only the successful instances are taken into account to calculate the average values, and these values are reported in brackets in Tables 1 and 2. The missing entries in these tables correspond to cases in which no instance could be solved in the required time limit. When the number of successful instances is lower than 10, this number is reported in Table 3 below, instead of the regular entry.

In the first batch of computational experiments, 15 series of 10 instances were generated. Each series is defined by a number of vertices ($n = 10, 30, 50$) and a density ($d = 0, 25, 50, 75, 100$). For each problem, the radius of the graph is computed,

TABLE 1
Dense graphs (average results over 10 *instances).*

n	Density	m	Radius	Maximum weight				
				$6 \times$ radius $- 1$	Tree	$\lceil C \rceil$	I	U
10	0 %	9	2.8	15.8	3.0	1.0	1.0	1.0
10	25 %	18	2.0	11.0	9.9	3.3	3.5	6.3
10	50 %	27	1.7	9.2	7.6	2.7	3.5	4.3
10	75 %	36	1.2	6.2	5.6	2.2	3.0	3.3
10	100 %	45	1.0	5.0	5.0	1.0	1.0	1.0
30	0 %	29	4.9	28.4	3.0	1.0	1.0	1.0
30	25 %	130	2.0	11.0	11.0	5.4	(6.0)	10.4
30	50 %	232	2.0	11.0	10.9	4.8	(5.0)	7.6
30	75 %	333	1.4	7.4	6.8	2.8	(3.7)	4.1
30	100 %	435	1.0	5.0	5.0	1.0	1.0	1.0
50	0 %	49	5.7	33.2	3.0	1.0	1.0	1.0
50	25 %	343	2.0	11.0	11.0	6.1	-	11.3
50	50 %	637	2.0	11.0	11.0	5.3	-	(8.5)
50	75 %	931	1.8	9.8	8.6	(3.8)	-	-
50	100 %	1225	1.0	5.0	5.0	1.0	1.0	1.0

TABLE 2
Sparse graphs (average results over 10 *instances).*

n	Density	m	Radius	Maximum weight				
				$6 \times$ radius $- 1$	Tree	$\lceil C \rceil$	I	U
10	0 %	9	2.8	15.8	3.0	1.0	1.0	1.0
10	16 %	15	2.1	11.6	9.4	3.2	3.3	5.5
10	30 %	20	2.0	11.0	9.9	3.5	3.5	5.7
10	58 %	30	1.3	6.8	6.1	2.3	3.2	3.5
30	0 %	29	4.9	28.4	3.0	1.0	1.0	1.0
30	3 %	45	3.6	20.6	19.3	6.3	6.5	14.4
30	7 %	60	3.0	17.0	16.2	5.8	(6.2)	14.4
30	15 %	90	2.5	14.0	13.0	5.6	(6.2)	11.2
50	0 %	49	5.7	33.2	3.0	1.0	1.0	1.0
50	2 %	75	4.4	25.4	20.9	7.0	7.0	17.7
50	4 %	100	3.7	21.2	19.3	6.8	(6.9)	17.7
50	8 %	150	3.0	17.0	16.7	6.8	-	15.2

a central vertex is selected, and a breadth-first search tree is built. According to the construction scheme of section 2, a weight is then assigned to each edge of the graph. The computational results displayed in the tables are average results over the 10 instances. The value of $6 \times \text{radius} - 1$ is also provided as a very basic bound, to be compared with the four other values.

The bound provided by the tree heuristic is better than $6 \times \text{radius} - 1$, but it is also much weaker than the values provided by the linear programs. It seems, in view of the few integer programs we were able to solve in less than an hour, that the lower bound C is rather close to the optimal integer value for the sparse graphs and starts to deteriorate when the density increases up to a relative gap of 30% in the worst cases. On the contrary, the upper bound U behaves quite badly for the sparse graphs with a relative gap reaching such values as 150% but becomes much tighter when the density increases, beating in absolute value the lower bound C. However, the combination of these two bounds allows us to obtain a good approximation of the optimal value.

Regarding the density of the graphs, the maximum value of the weight seems to

Table 3
CPU time (in seconds).

n	Density	m	Tree	$\lceil C \rceil$	I	U
10	0 %	9	0.01	0.02	0.02	0.02
10	25 %	18	0.01	0.21	0.39	0.23
10	50 %	27	0.02	0.30	1.19	0.37
10	75 %	36	0.02	0.37	1.05	0.57
10	100 %	45	0.01	0.02	0.01	0.02
30	0 %	29	0.01	0.73	0.73	0.73
30	25 %	130	0.12	31.25	(4)	38.55
30	50 %	232	0.16	150.18	(3)	171.53
30	75 %	333	0.21	137.52	(9)	239.52
30	100 %	435	0.22	0.66	0.67	0.64
50	0 %	49	0.24	5.25	5.26	5.25
50	25 %	343	0.35	606.21	(0)	821.03
50	50 %	637	0.54	2973.37	(0)	(4)
50	75 %	931	0.77	(2)	(0)	(0)
50	100 %	1225	0.93	4.76	4.74	4.73

increase rapidly from the value 1 for the tree and then slowly decreases back to the value 1, also optimal for the complete graph. The maximum value of the weight is also related to the size of the graph since the values observed are larger for graphs with more nodes for a given level of density.

In the second batch of computational experiments, 12 series of 10 instances were generated. Each series is again defined by a number of vertices ($n = 10, 30, 50$) and a number of edges proportional to the number of nodes with a coefficient of 1, 1.5, 2, and 3. The graphs generated in this batch are hence somewhat sparser than the ones of the first batch.

The observations made on the first batch are confirmed with the second batch. The peak point in the evolution of the maximum weights seems to be located somewhere in the sparse graphs area, apparently very close to the trees.

Finally, in the third batch of computational experiments, the computing time of the various approaches are compared. The average computing time in seconds, over the 10 instances, is reported in Table 3. When some instances required more than the one hour time limit, the number reported in brackets is the number of successful instances, i.e., the number of instances solved in less than an hour.

Although the tree heuristic and the linear programming approaches are definitively not designed for the same purpose, the results displayed in Table 3 show how much faster this simple heuristic can be. It almost never takes more than a second to build the breadth-search tree and obtain the bound. The more elaborate approaches based on various linear programs address a more difficult and general problem, since they allow us to find weights corresponding to any given pattern of routing paths (if such weights exist). As an obvious drawback, these methods are much more time-consuming and the computing time required even increases rapidly with the size of the problems. Even more so, for many problems, the exact solution (that is, the minimum maximal weight of a set of integer weights) cannot be reached in a reasonable amount of time (with the proposed approach).

Finally, for some instances, we observed that $I > C + 1$. This implies that we cannot hope (in general) to deduce integer weights only by rounding the weights given by LP_4.

5. Conclusion and open problems. In this paper, we solved some important problems related to Internet networks. The traffic demands are carried through shortest paths in sense of a set of administrative link weights. We showed that we can simultaneously use all the network resources (links), avoid routing ambiguities (a unique shortest path for each demand), and use a set of integer weights which are strictly lower than 6 times the radius of the graph (and hence can be easily encoded on routers). Then, we studied the problem of computing a set of weights that is compatible with a given set of routing paths. The routing paths can be the output of a dimensioning procedure [4]. In other cases, they are imposed by technical and topological constraints. We gave some necessary conditions that must be satisfied by the routing paths to find compatible weights. Linear programs enabling the solution of this problem were presented. Focus was on algorithms that can be used to compute integer weights. We also found many important particular graphs such as cycles, cacti, etc., for which the suboptimality condition is sufficient to find weights.

Many problems connected with those solved in this paper are still open. First, is it possible to have a general upper bound on the integer weights that is better than six times the radius of the graph? Second, is there any general "simple" necessary and sufficient condition that must be satisfied by the routing paths to guarantee the existence of a set of compatible weights?

Computing a set of real weights compatible with some routing paths is a polynomial problem, but what is the complexity of this problem when the weights have to be integers and as small as possible?

Another interesting problem that has not been addressed in this paper can be stated as follows: given a graph $G = (V, E)$, can we find a minimal (in the cardinal sense) set of values $W = \{w_1, w_2, \ldots\}$ that generates all the possible shortest routing paths on G? As we know that the set of weights that are compatible with a set of routing paths ϕ is either empty or is a nonempty polyhedron P_ϕ, the problem described above is equivalent to finding a minimal set W such that $P_\phi \cap \big(W^{|E|} = W \times W \times \cdots \times W\big) \neq \emptyset$ for any set of routing paths ϕ for which $P_\phi \neq \emptyset$.

REFERENCES

[1] C. Barnhart, C. Hane, and P. Vance, *Using branch-and-price to solve origin-destination integer multicommodity flow problems*, Oper. Res., 2 (2000), pp. 318–326.

[2] W. Ben-Ameur, *Constrained length connectivity and survivable networks*, Networks, 36 (2000), pp. 17–33.

[3] W. Ben-Ameur and E. Gourdin, *Exact Algorithms for Internet Dimensioning*, Technical report DE/DAC/OAT/30.99, France Télécom R&D, Issy-les-Moulineaux, France, 1999 (in French).

[4] W. Ben-Ameur, E. Gourdin, B. Liau, and N. Michel, *Dimensioning of Internet networks*, in Proceedings of the Second International Workshop on the Design of Reliable Communication Networks, Munich, 2000, pp. 56–61.

[5] W. Ben-Ameur, E. Gourdin, B. Liau, and N. Michel, *Routing strategies for IP networks*, in Telektronikk Magazine, 2/3 (2001), pp. 145–158.

[6] C. Berge, *Graphes et Hypergraphes*, Dunod, Paris, 1970.

[7] B. Bollabás, *Random Graphs*, Academic Press, London, 1985.

[8] D. Burton and Ph.L. Toint, *On an instance of the inverse shortest paths problem*, Math. Program., 53 (1992), pp. 45–61.

[9] K. L. Calvert, M. B. Doar, and E. W. Zegura, *Modeling Internet topology*, IEEE Communications Magazine, 35 (1997), pp. 160–163.

[10] ILOG, Inc., *Using the CPLEX Linear Optimizer*, Incline Village, NV,

[11] E. Crawley, R. Nair, R. Rajagopalan, and H. Sandick, *A Framework for QoS-Based Routing in the Internet*, IETF Request for comments 2386, 1998.

[12] R. Diestel, *Graph Theory*, Springer, 1997.

[13] Y. Dinitz, N. Garg, and M. X. Goemans, *On the single-source unsplittable flow problem*, Combinatorica, 19 (1999), pp. 1-25.

[14] S. E. Dreyfus, *An appraisal of some shortest-path algorithms*, ORSA J. Comput., 17 (1969), pp. 395–412.

[15] M. Faloutsos, P. Faloutsos, and C. Faloutsos, *On power-law relationships of the internet topology*, in Proceedings of SIGCOMM '99, ACM, New York, 1999.

[16] A. Farago, A. Szentesi, and B. Szviatovszki, *Allocation of administrative weights in PNNI*, in Proceedings of Networks '98, Italy, 1998, pp. 621–625.

[17] B. Fortz and M. Thorup, *Internet traffic engineering by optimizing OSPF weights*, in Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies, Vol. 2, IEEE Press, Piscataway, NJ, 2000, pp. 519–528.

[18] ATM Forum, *Private network-network interface, specification version* 1.0, af-pnni-0055.00, 1996.

[19] P. Gajowniczek, M. Pióro, M. Szentesi, and A. Harmatos, *Solving an OSPF routing problem with simulated allocation*, in Proceedings of the First Polish-German Teletraffic Symposium, Dresden, 1998.

[20] J. Geffard, *A* 0-1 *model for singly routed traffic in telecommunications*, Annals Telecommunications, 56 (2001), pp. 140–149.

[21] M. Gondran and M. Minoux, *Graphes et algorithmes*, Eyrolles, Paris, 1995.

[22] M. Grötschel, L. Lovász, and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–197.

[23] C. Hedrick, *Routing Information Protocol*, IETF Request for comments 1058, 1988.

[24] C. Huitema, *Le routage dans l'Internet*, Eyrolles, 1995.

[25] N. Katoh, T. Ibaraki, and H. Mine, *An efficient algorithm for K shortest simple paths*, Networks, 12 (1982), pp. 411–427.

[26] D. Lorenz and A. Orda, *QoS routing in networks with uncertain parameters*, IEEE/ACM Transactions on Networking, 6 (1998), pp. 768–778.

[27] D. Lorenz, A. Orda, D. Raz, and Y. Shavitt, *How good can IP routing be?*, Technical report 2001-17, DIMACS, Rutgers University, Piscataway, NJ, 2001.

[28] K. Mehlhon, S. Näher, M. Seel, and C. Uhrig, *Efficient Data Types and Algorithms*, version 3.7., LEDA Software GmbH, Saarbrücken, Germany.

[29] M. Minoux, *Programmation Mathématique-Théorie et Algorithmes*, Dunod, Paris, 1983.

[30] J. Moy, *OSPF* Version 2, IETF Request for comments 1583, 1994.

[31] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*, John Wiley, New York, 1988.

[32] D. Oran, *OSI IS-IS Intra-Domain Routing Protocol*, IETF Request for comments 1142, 1990.

[33] K. Park, S. Kang, and S. Park, *An integer programming approach to the bandwidth packing problem*, Management Science, 42 (1996), pp. 1277–1291.

[34] L. Peterson and B. Davie, *Computer Networks, A Systems Approach*, Morgan Kaufmann, San Francisco, CA, 2000.

[35] A. Schrijver, P. Seymour, and P. Winkler, *The ring loading problem*, SIAM J. Discrete Math., 11 (1998), pp. 1–14.

[36] Z. Wang and J. Crowcroft, *Quality-of-service routing for supporting multimedia applications*, IEEE J. Selected Areas Communications, 14 (1996), pp. 1228–1234.

[37] Y. Wang, Z. Wang, and L. Zhang, *Internet traffic engineering without full mesh overlaying*, in Proceedings of the Twentieth Annual Joint Conference of the IEEE Computer and Communication Societies, Vol. 1, IEEE Press, Piscataway, NJ, 2001, pp. 565–571.

[38] B. M. Waxman, *Routing of multipoint connections*, IEEE J. Selected Areas Communications, 6 (1988), pp. 1617–1622.

[39] J. Y. Yen, *Finding the K shortest loopless paths in a network*, Management Science, 17 (1971), pp. 712–716.

[40] J. Zinky, G. Vichniac, and A. Khanna, *Performance of the revised routing metric in the ARPANET and MILNET*, In proceedings of the Military Communications Conference, Vol. 1, IEEE, 1989, pp. 219–224.

SIAM J. Discrete Math.
Vol. 17, No. 1, pp. 50–71

ON THE HOMOLOGY OF THE h, k-EQUAL DOWLING LATTICE*

ERIC GOTTLIEB†

Dedicated to my sister Karla and her husband Josh in honor of the birth of their first child, Xavier Eli

Abstract. We define a Dowling lattice generalization of the k-equal partition lattice and the h, k-equal signed partition lattice. We use the theory of lexicographical shellability to show that it has the homotopy type of a wedge of spheres, to describe its Betti numbers, and to give a basis for its homology in terms of labelled trees. For cyclic groups, the h, k-equal Dowling lattice arises as the lattice of intersections of a complex subspace arrangement. We use Whitney homology to compute the cohomology of the complement of this arrangement. Our results generalize and unify previous research.

Key words. Dowling lattice, lexicographical shellability, poset, homology, subspace arrangement

AMS subject classifications. 05E25, 06B99

DOI. 10.1137/S0895480103422482

1. Background.

Extensive work has been done on the topology of the partition lattice Π_n, the signed partition lattice $\overline{\Pi}_n$, and the Dowling lattice $Q_n(G)$; see [Ba], [Be], [Go-Wa], [Ha], [St], and [Wa2], for example. Restricted block size partition posets have also been studied, as in [Bj-Lo-Ya], [Bj-Sa], sections 6 and 7 of [Bj-Wa3], [Bj-We], [Bw], [Ca-Ha-Ro], [Ha-Wa], [Li], [Su1], [Su2], [Su3], [Su-Wa], [Wa1], and [Wa3].

The posets Π_n and $\overline{\Pi}_n$ are both special cases of $Q_n(G)$. We will examine the topology of $Q_n^{h,k}(G)$, a restricted block size Dowling lattice. This poset has as special cases the k-equal partition lattice first studied in [Bj-Lo-Ya] and the h, k-equal signed partition lattice of [Bj-Sa].

The posets Π_n and $\overline{\Pi}_n$ arise as intersection lattices of real hyperplane arrangements. Specifically, Π_n is the intersection lattice of the hyperplane arrangement

$$\mathcal{A}_n = \{x_i = x_j \mid 1 \leq i < j \leq n\}$$

and $\overline{\Pi}_n$ is the intersection lattice of the hyperplane arrangement

$$\mathcal{B}_n = \{x_i = \pm x_j \mid 1 \leq i < j \leq n\} \cup \{x_i = 0 \mid 1 \leq i \leq n\}.$$

Let C_m denote the group generated by ω, a primitive mth root of unity. $Q_n(C_m)$ is the intersection lattice of the complex hyperplane arrangement

$$\mathcal{B}_{n,m} = \{x_i = \omega^p x_j \mid 0 \leq p < m, 1 \leq i < j \leq n\} \cup \{x_i = 0 \mid 1 \leq i \leq n\},$$

which is known as the Dowling arrangement. The Orlik–Solomon formula [Or-So] describes the cohomology of the complement of a hyperplane arrangement in terms of the Whitney homology of the arrangement's intersection lattice. One can determine

*Received by the editors February 7, 2003; accepted for publication March 28, 2003; published electronically September 17, 2003. This research was supported in part by a faculty development grant from Rhodes College.

http://www.siam.org/journals/sidma/17-1/42248.html

†Department of Mathematics and Computer Science, Rhodes College, Memphis, TN 38112-1690 (gottlieb@rhodes.edu).

the Whitney homologies of Π_n, $\overline{\Pi}_n$, and $Q_n(G)$ once their homologies are known. Thus results about the topologies of Π_n, $\overline{\Pi}_n$, and $Q_n(C_m)$ can be used to understand the topology of complements of certain hyperplane arrangements.

The k-equal partition lattice Π_n^k is the sublattice of Π_n consisting of those $\pi \in \Pi_n$ such that each block B of π satisfies $|B| = 1$ or $|B| \geq k$. It first appeared in the work of Björner, Lovász, and Yao [Bj-Lo-Ya], where they studied the computational complexity of the following question: given a string of n numbers, are any k of them the same? Their results depended on computing the Möbius number (i.e., the reduced Euler characteristic) of Π_n^k, which appeared as the intersection lattice of the k-equal subspace arrangement

$$\mathcal{A}_n^k = \{x_{i_1} = \cdots = x_{i_k} \mid 1 \leq i_1 < \cdots < i_k \leq n\}.$$

Björner and Welker [Bj-We] computed the homotopy type of Π_n^k and used the Goresky–MacPherson formula[1] [Go-Ma] to determine the cohomology of the complement of $\mathcal{A}_n^k$ and Betti numbers for this space.

Björner and Wachs [Bj], [Bj-Wa1], [Bj-Wa2], [Bj-Wa3], [Bj-Wa4] have adapted and generalized shellability, a powerful tool from polytope theory, to a tool for poset topology called lexicographical shellability. We will be using a version of lexicographical shellability called EL-shellability. We now describe some of the terminology and results of this theory.

Let x and y be elements of a poset P. When $x < y$ and there is no $z \in P$ such that $x < z < y$ we say that x *covers* y. In this case we write $x \lessdot y$ and refer to $x \lessdot y$ as a *cover* of P. The *covering relation* of P is the set of covers of P, which we denote by $CR(P)$. We say that a chain of P is *maximal* if there is no other chain of P that contains it. Let $x_1 < x_2$ in P. We say that a chain of $[x_1, x_2]$ is *saturated* in P if it is maximal in $[x_1, x_2]$.

Given a totally ordered set Γ and a map $\lambda : CR(P) \to \Gamma$ we refer to $\lambda(x \lessdot y)$ as the *label* of $x \lessdot y$ and to λ as a *labelling* of P. We associate with a saturated chain $c : x_0 \lessdot \cdots \lessdot x_p$ the *label sequence* $\lambda(x_0 \lessdot x_1), \ldots, \lambda(x_{p-1} \lessdot x_p)$, which we denote by $\lambda(c)$. A saturated chain c and its label sequence $\lambda(c)$ are called λ-*increasing* if $\lambda(c)$ is strictly increasing and λ-*decreasing* if $\lambda(c)$ is weakly decreasing. We use the terms *increasing* and *decreasing* if λ is understood. If c' is another saturated chain of P, we write $c < c'$ if $\lambda(c) < \lambda(c')$ lexicographically. λ is called an *EL-shelling* of P if every interval of P has a unique increasing maximal chain that precedes every other maximal chain of that interval.

A poset P is *bounded* if it contains a minimum element and a maximum element and these elements are distinct. We will often denote these elements by $\hat{0}$ and $\hat{1}$, respectively. If P is a bounded finite poset with an EL-shelling, then P is said to be *EL-shellable*. If c is a maximal chain of P, then let $\hat{c}$ denote the maximal chain $c \setminus \{\hat{0}, \hat{1}\}$ of $P \setminus \{\hat{0}, \hat{1}\}$.

Let P be a finite bounded poset. The chains of $P \setminus \{\hat{0}, \hat{1}\}$ form an (abstract) simplicial complex called the *order complex* of P. We denote this complex by $\Delta(P)$. The field of poset topology is concerned with the topological properties of order complexes of posets. If a topological statement is true of $\Delta(P)$, then we say that the same statement is true of P. We denote the ith reduced homology and cohomology of $\Delta(P)$ over the integers by $\tilde{H}_i(P)$ and $\tilde{H}^i(P)$, respectively.

[1] The Goresky–MacPherson formula is a partial generalization of the Orlik–Solomon formula. It describes the cohomology of the complement of a subspace arrangement. However, it does not take group actions into account. An equivariant version is given by Sundaram and Welker [Su-We].

The following theorem is one of the main results of the theory of EL-shellability.

Theorem 1.1 (see [Bj-Wa3, Theorem 5.9]). *An EL-shellable poset P has the homotopy type of a wedge of spheres. Furthermore, for any EL-shelling λ of P,*

- $\tilde{H}_i(P) \cong \mathbf{Z}^{\#\text{ of }\lambda\text{- decreasing chains of }P\text{ of length i+2}}$,
- *the set $\{\hat{c} \mid c$ is a λ-decreasing chain of P of length $i+2\}$ forms a basis for $\tilde{H}^i(P)$.*

Some EL-shellings for Π_n are given in [Bj] and in [Wa1]. An EL-shelling for $Q_n(G)$ is given in [Go-Wa]. [Bj-Wa3] contains the EL-shelling for Π_n^k defined below. We denote a partition $\{B_1, \ldots, B_j\}$ of a finite set S by $B_1/\cdots/B_j$. A cover $x \lessdot y$ in Π_n corresponds to a merge of blocks of x. That is, if $x = B_1/\cdots/B_j$, then y is of the form $B_1/\cdots/B_m \cup B_p/\cdots/B_j$ for some m and p, $1 \leq m < p \leq j$. We denote such a merge by $B_m/B_p \lessdot B_m \cup B_p$.

The situation for Π_n^k is similar but there are three kinds of merges. The first is a merge of two nonsingleton blocks B and B'. We denote this type of merge by $B/B' \lessdot B \cup B'$. The second is a merge of a singleton block and a nonsingleton block. We denote this type of merge by $\{a\}/B \lessdot \{a\} \cup B$, where B is a nonsingleton. The third is a merge of k singleton blocks and is denoted by $\{a_1\}/\cdots/\{a_k\} \lessdot \{a_1, \ldots, a_k\}$.

Theorem 1.2 (see [Bj-Wa3, Theorem 6.1]). *The labelling λ of Π_n^k defined by*

$$\lambda(x \lessdot y) = \begin{cases} (1, \max\,(B \cup B')), & B/B' \lessdot B \cup B', \\ (2, a), & \{a\}/B \lessdot \{a\} \cup B, \\ (2, \max\,\{a_1, \ldots, a_k\}), & \{a_1\}/\cdots/\{a_k\} \lessdot \{a_1, \ldots, a_k\} \end{cases} \tag{1.1}$$

is an EL-shelling for Π_n^k.

Here and throughout this paper, pairs used to label covers for the purposes of a shelling are ordered lexicographically.

Björner and Sagan [Bj-Sa] defined the h,k-equal signed partition poset $\overline{\Pi}_n^{h,k}$ to be the subposet of $\overline{\Pi}_n$ consisting of those signed partitions whose nonzero blocks have size 1 or size at least k, and whose zero blocks have size 1 or size at least $h+1$. They showed that the h,k-equal signed partition lattice is isomorphic to the intersection lattice of the h,k-equal subspace arrangement

$$\mathcal{B}_n^{h,k} = \{\pm x_{i_1} = \cdots = \pm x_{i_k}\} \cup \{x_{j_1} = \cdots = x_{j_h} = 0\},$$

where $1 \leq i_1 < \cdots < i_k \leq n$ and $1 \leq j_1 < \cdots < j_h \leq n$. They gave an EL-shelling of $\overline{\Pi}_n^{h,k}$, described the Betti numbers of $\overline{\Pi}_n^{h,k}$, and used them to obtain the ranks of the cohomology of the complement of $\mathcal{B}_n^{h,k}$ when $k > 2$. We postpone the description of their EL-shelling until we develop notation for Dowling lattices.

In this paper, we define a subposet $Q_n^{h,k}(G)$ of $Q_n(G)$ which generalizes Π_n^k and $\overline{\Pi}_n^{h,k}$. We refer to $Q_n^{h,k}(G)$ as the *h,k-equal Dowling lattice.* We define a complex subspace arrangement $\mathcal{B}_{n,m}^{h,k}$ which generalizes $\mathcal{A}_n^k$ and $\mathcal{B}_n^{h,k}$. We refer to $\mathcal{B}_{n,m}^{h,k}$ as the *h,k-equal Dowling arrangement.* We show that $Q_n^{h,k}(C_m)$ is the lattice of intersections of $\mathcal{B}_{n,m}^{h,k}$. We give two closely related EL-shellings for $Q_n^{h,k}(G)$. We use the EL-shelling to describe bases for the (co)homology of $Q_n^{h,k}(G)$, to give an expression for the Betti numbers of $Q_n^{h,k}(G)$, and to describe the ranks of the cohomology of the complement of $\mathcal{B}_{n,m}^{h,k}$.

2. The h,k-equal Dowling lattice $Q_n^{h,k}(G)$.

We begin by reviewing the definition of the Dowling lattice. Let G denote a finite group with identity e. Let $\pi = B_0/B_1/\cdots/B_j$ be a partition of $\{0, 1, \ldots, n\}$. Throughout this paper we will assume that $0 \in B_0$. Let $\gamma : B_1 \cup \cdots \cup B_j \to G$ and denote the restriction of γ to B_p by γ_{B_p}. If we think of γ formally (that is, as a subset of $(B_1 \cup \cdots \cup B_j) \times G$), then π induces the partition $B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$ of $B_0 \cup \gamma$. We refer to $B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$ as a *G-prepartition* of $\{0, 1, \ldots, n\}$. We refer to B_0 as the *zero block* of π and of $B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$. For $p > 0$ we say that B_p is a *nonzero block* of π and that γ_{B_p} is a nonzero block of $B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$.

For example, setting $\pi = \{0, 3, 6\}/\{1, 2, 7\}/\{4, 5\}$ and

$$\gamma = \{(1, -1), (2, 1), (4, 1), (5, -1), (7, 1)\}$$

gives the C_2-prepartition

$$\{0, 3, 6\}/\{(1, -1), (2, 1), (7, 1)\}/\{(4, 1), (5, -1)\}$$

of $\{0, 1, \ldots, 7\}$.

When representing particular partitions of $\{0, 1, \ldots, n\}$ we will drop curly brackets and commas. Thus the partition of $\{0, 1, \ldots, 7\}$ in the example will be written 036/127/45. We also drop curly brackets and commas when discussing G-prepartitions. Additionally, we drop parentheses and put the second coordinate above the corresponding first coordinate unless the second coordinate is the identity, in which case we will put nothing above the first coordinate. When our examples involve C_2-prepartitions (as they usually will), and when the second coordinate is -1, we will simply bar the first coordinate. Thus the example above will be written $036/\bar{1}27/4\bar{5}$.

There is a right action of G on $\{1, \ldots, n\} \times G$ given by $(i, h) \cdot g = (i, hg)$. This action extends to subsets S of $\{1, \ldots, n\} \times G$ by $S \cdot g = \{x \cdot g \mid x \in S\}$. We say that two subsets S_1 and S_2 of $\{1, \ldots, n\} \times G$ are *equivalent* if there exists $g \in G$ so that $S_1 = S_2 \cdot g$. This is an equivalence relation on the power set of $\{1, \ldots, n\} \times G$. We will usually denote the equivalence class of S by S, resorting to $\overline{S}$ only when necessary.

Two G-prepartitions are *equivalent* if their zero blocks are equal and there is a one-to-one correspondence between their nonzero blocks such that corresponding nonzero blocks are equivalent. This is an equivalence relation on G-prepartitions. A *G-partition* of $\{0, 1, \ldots, n\}$ is the equivalence class of a G-prepartition of $\{0, 1, \ldots, n\}$. As above, we will denote the equivalence class of a G-prepartition $B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$ by $\overline{B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}}$ when useful, but normally we will omit the overbar.

The equivalence class of the C_2-prepartition from the example is

$$\{036/\bar{1}27/4\bar{5}, 036/1\overline{27}/4\bar{5}, 036/\bar{1}27/\bar{4}5, 036/1\overline{27}/\bar{4}5\}.$$

We will usually represent an equivalence class with the member in which the smallest element of each nonzero block is labelled with the identity. We will refer to this representative as the canonical representative. In this case, the canonical representative is $036/1\overline{27}/4\bar{5}$.

The equivalence class of a nonzero block of a G-prepartition will be called a *nonzero G-block* of the corresponding G-partition. We say that a nonzero G-block γ_{B_p} is a *singleton* if $|B_p| = 1$ and a *nonsingleton* if $|B_p| > 1$. We refer to the zero block of a G-prepartition as the *zero G-block* of the corresponding G-partition. A *G-block* of a G-partition is either a zero G-block or a nonzero G-block of the G-partition.

The Dowling lattice $Q_n(G)$ is the set of G-partitions of $\{0, 1, \ldots, n\}$ with partial order determined by the covering relation described below. An element y covers $B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$ if y results from merging two G-blocks of $B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$. There are two ways to do this. One way involves B_0; the other does not.

B_0 can be merged with a nonzero G-block γ_{B_p} of $B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$ yielding the G-partition $y = B_0 \cup B_p/\gamma_{B_1}/\cdots/\gamma_{B_j}$. There are two merges of this type for the C_2-partition from the example. They give the covers

$$036/1\overline{27}/4\overline{5} \lessdot 012367/4\overline{5} \quad \text{and} \quad 036/1\overline{27}/4\overline{5} \lessdot 03456/1\overline{27}.$$

We will denote merges that involve B_0 by $B_0/\gamma_{B_p} \lessdot B_0 \cup B_p$.

There are $|G|$ covers of $B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$ in which two distinct nonzero G-blocks γ_{B_p} and γ_{B_m} are merged, each of which is of the form

$$B_0/\gamma_{B_1}/\cdots/\gamma_{B_p} \cup (\gamma_{B_m} \cdot g)/\cdots/\gamma_{B_j}$$

for some $g \in G$. The covers that result from merging the nonzero C_2-blocks of the C_2-partition in the example are

$$036/1\overline{27}/4\overline{5} \lessdot 036/1\overline{2}4\overline{57} \quad \text{and} \quad 036/1\overline{27}/4\overline{5} \lessdot 036/1\overline{24}5\overline{7}.$$

Consider the merge of two nonzero G-blocks γ_{B_p} and γ_{B_m} from $B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$ into the G-block $\gamma_{B_p} \cup (\gamma_{B_m} \cdot g)$. Define $\delta : B_1 \cup \cdots \cup B_j \to G$ by

$$\delta(i) = \begin{cases} \gamma(i) \cdot g, & i \in B_m, \\ \gamma(i), & \text{else}, \end{cases}$$

and observe that $\overline{\gamma_{B_i}} = \overline{\delta_{B_i}}$ for $i = 1, \ldots, j$. Also, $\overline{\gamma_{B_p} \cup (\gamma_{B_m} \cdot g)} = \overline{\delta_{B_p} \cup \delta_{B_m}}$. In other words, by choosing a suitable equivalence class representative, we can suppress the group multiplier in the merge. Without loss of generality, we may assume that $B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$ is that representative. Thus our notation for such a merge need not include the group multiplier. We denote a merge of this type by $\gamma_{B_p}/\gamma_{B_m} \lessdot \gamma_{B_p \cup B_m}$.

The smallest element of $Q_n(G)$ is $\hat{0} = \{0\}/\gamma_{\{1\}}/\cdots/\gamma_{\{n\}}$, where γ is arbitrary. The largest element of $Q_n(G)$ is $\hat{1} = \{0, 1, \ldots n\}$. The two most familiar instances of the Dowling lattice are the partition lattice Π_{n+1}, which is isomorphic to $Q_n((e))$, and the signed partition lattice $\overline{\Pi}_n$, which is isomorphic to $Q_n(C_2)$.

We assume throughout this paper that n, h, and k are integers satisfying $1 \le h < k$ and $n \ge h$ and $n \ge 2$. We define $Q_n^{h,k}(G)$ to be the subposet of $Q_n(G)$ consisting of those G-partitions for which the zero G-block satisfies $|B_0| = 1$ or $|B_0| \ge h + 1$ and each nonzero G-block γ_B satisfies $|B| = 1$ or $|B| \ge k$.

- $Q_n^{h,k}(G)$ is not pure when $h > 1$ and $n \ge h + k$ nor when $k > 2$ and $n \ge 2k$.
- $Q_n^{1,2}(G) = Q_n(G)$.
- $Q_n^{k-1,k}((e)) = \Pi_{n+1}^k$.
- $Q_n^{1,2}(C_2) = \overline{\Pi}_n$.
- $Q_n^{h,k}(C_2) = \overline{\Pi}_n^{h,k}$.

Note that if $n < h$, then $Q_n^{h,k}(G)$ consists of a single element so is not bounded. The assumption that $h < k$ makes $Q_n^{h,k}(G)$ a lattice. Without this assumption, the conclusion fails in some cases. When $h < k - 1$ the poset $Q_n^{h,k}((e))$ is a sublattice of Π_{n+1} that to our knowledge has not been studied previously.

Every covering relation in $Q_n^{h,k}(G)$ corresponds to a merge of exactly one of the six following types.

- A zero-nonsingleton (Z-NS) merge is a merge between the zero G-block B_0 and a nonzero nonsingleton G-block γ_B. This merge is represented by $B_0/\gamma_B \lessdot B_0 \cup B$. There is no restriction on $|B_0|$ other than $|B_0| = 1$ or $|B_0| \ge h + 1$.

- An h-merge is a merge of $\{0\}$ with the h nonzero singleton G-blocks $\gamma_{\{a_1\}}, \dots,$ $\gamma_{\{a_h\}}$. It is represented by $\{0\}/\gamma_{\{a_1\}}/\cdots/\gamma_{\{a_h\}} \lessdot \{0, a_1, \dots, a_h\}$.
- A zero-singleton (Z-S) merge is a merge between the zero G-block B_0 and a nonzero singleton G-block $\gamma_{\{a\}}$ where $|B_0| \geq h+1$. It is represented by $B_0/\gamma_{\{a\}} \lessdot B_0 \cup \{a\}$.
- A nonsingleton-nonsingleton (NS-NS) merge is a merge between two nonzero nonsingleton G-blocks γ_B and $\gamma_{B'}$. It is represented by $\gamma_B/\gamma_{B'} \lessdot \gamma_{B\cup B'}$.
- A k-merge is a merge of k nonzero singleton G-blocks $\gamma_{\{a_1\}}, \dots, \gamma_{\{a_k\}}$. It is represented by $\gamma_{\{a_1\}}/\cdots/\gamma_{\{a_k\}} \lessdot \gamma_{\{a_1,\dots,a_k\}}$.
- A singleton-nonsingleton (S-NS) merge is a merge between a nonzero singleton G-block $\gamma_{\{a\}}$ and a nonzero nonsingleton G-block γ_B. It is represented by $\gamma_{\{a\}}/\gamma_B \lessdot \gamma_{\{a\}\cup B}$.

We now describe Björner and Sagan's EL-shelling. Since $\overline{\Pi}_n^{h,k} = Q_n^{h,k}(C_2)$ we use the notation we have developed for Dowling lattices.

THEOREM 2.1 (see [Bj-Sa, Theorem 4.4]). *Suppose* $1 \leq h < k$ *and* $n \geq h$. *The labelling* λ *of* $\overline{\Pi}_n^{h,k}$ *defined by*

$$\lambda(x \lessdot y) = \begin{cases} (1, \max B), & B_0/\gamma_B \lessdot B_0 \cup B, \\ (2, \max_i a_i), & \{0\}/\gamma_{\{a_1\}}/\cdots/\gamma_{\{a_h\}} \lessdot \{0, a_1, \dots, a_h\}, \\ (2, a), & B_0/\gamma_{\{a\}} \lessdot B_0 \cup \{a\}, \\ (3, \max(B \cup B')), & \gamma_B/\gamma_{B'} \lessdot \gamma_{B\cup B'}, \\ (4, \max_i a_i), & \gamma_{\{a_1\}}/\cdots/\gamma_{\{a_k\}} \lessdot \gamma_{\{a_1,\dots,a_k\}}, \\ (4, a), & \gamma_{\{a\}}/\gamma_B \lessdot \gamma_{\{a\}\cup B} \end{cases}$$

is an EL-shelling.

3. Two EL-shellings for $Q_n^{h,k}(G)$. We now give an EL-shelling for $Q_n^{h,k}(G)$ that is based on that of Björner and Wachs (our Theorem 1.2) for Π_n^k. Setting $h = k-1$ and $G = (e)$ gives a new EL-shelling for Π_{n+1}^k because of the special treatment given to the letter 0. If $k = 2$, we also get a new shelling of Π_{n+1}. Setting $G = C_2$ gives a new EL-shelling for $\overline{\Pi}_n^{h,k}$ that is related to but simpler than Björner and Sagan's EL-shelling (our Theorem 2.1).

THEOREM 3.1. *For* $1 \leq h < k$ *and* $n \geq h$, *the labelling* λ *of* $Q_n^{h,k}(G)$ *defined by*

$$\lambda(x \lessdot y) = \begin{cases} (1, \max B), & B_0/\gamma_B \lessdot B_0 \cup B, \\ (2, \max_i a_i), & \{0\}/\gamma_{\{a_1\}}/\cdots/\gamma_{\{a_h\}} \lessdot \{0, a_1, \dots, a_h\}, \\ (2, a), & B_0/\gamma_{\{a\}} \lessdot B_0 \cup \{a\}, \\ (1, \max(B \cup B')), & \gamma_B/\gamma_{B'} \lessdot \gamma_{B\cup B'}, \\ (2, \max_i a_i), & \gamma_{\{a_1\}}/\cdots/\gamma_{\{a_k\}} \lessdot \gamma_{\{a_1,\dots,a_k\}}, \\ (2, a), & \gamma_{\{a\}}/\gamma_B \lessdot \gamma_{\{a\}\cup B} \end{cases}$$

is an EL-shelling.

Proof. Let $S \subseteq \{0, 1, \dots, n\}$ with $0 \in S$ and $|S| > h$. There is a natural way to define a poset $Q_S^{h,k}(G)$ and a labelling $\lambda_S : CR(Q_S^{h,k}(G)) \to \{1, 2\} \times (S \setminus \{0\})$ so that $Q_{\{0,1,\dots,n\}}^{h,k}(G) = Q_n^{h,k}(G)$ and $\lambda_{\{0,1,\dots,n\}} = \lambda$. We will prove by induction on $|S|$ that λ_S is an EL-shelling.

For the base step, suppose $S = \{0, a_1, \dots, a_h\}$. Note that $Q_S(G)$ is the chain $\{0\}/\gamma_{\{a_1\}}/\cdots/\gamma_{\{a_h\}} \lessdot S$. Assigning any label to this cover gives a shelling of $Q_S(G)$, so the label $\lambda_S(\{0\}/\gamma_{\{a_1\}}/\cdots/\gamma_{\{a_h\}} \lessdot S) = (2, \max_i a_i)$ works.

For the induction step, suppose that $|S| > h+1$. We must show that every interval $[x, y]$ of $Q_S^{h,k}(G)$ has a unique increasing chain and that this chain is lexicographically first among the maximal chains of $[x, y]$.

First we suppose that $y \neq \hat{1}$ and let $y = B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$ where $j > 0$. The interval $[\hat{0}, y]$ is isomorphic to the product

$$Q_{B_0}^{h,k}(G) \times \Pi_{B_1}^k \times \cdots \times \Pi_{B_j}^k. \tag{3.1}$$

For any $T \subseteq \{1, \ldots, n\}$ there is a natural way to define a lattice Π_T^k and an EL-shelling $\kappa_T : CR(\Pi_T^k) \to \{1, 2\} \times T$ of Π_T^k so that $\Pi_{\{1,\ldots,n\}}^k$ is Π_n^k and $\kappa_{\{1,\ldots,n\}}$ is the EL-shelling of Theorem 1.2 for Π_n^k.

λ_S restricts[2] to λ_{B_0}. By induction, we may assume that λ_{B_0} is an EL-shelling of $Q_{B_0}^{h,k}(G)$. Also, λ_S restricts to κ_{B_p}, which is an EL-shelling of $\Pi_{B_p}^k$. The ranges of λ_{B_0}, $\kappa_{B_1}, \ldots, \kappa_{B_j}$ are pairwise disjoint and the total order on the range of λ_S is a shuffle of the total orders on the ranges of $\lambda_{B_0}, \kappa_{B_1}, \ldots, \kappa_{B_j}$. We can therefore apply Proposition 10.15 of [Bj-Wa4] to conclude that the labelling for (3.1) is an EL-shelling. This labelling is respected by the isomorphism from (3.1) to the restriction of λ_S to $[\hat{0}, y]$. The restriction of λ_S to $[\hat{0}, y]$ is therefore an EL-shelling of $[\hat{0}, y]$. It follows that for any $x < y$ there is a unique increasing chain in $[x, y]$ that lexicographically precedes all other maximal chains in $[x, y]$.

Now suppose that the interval under consideration is of the form $[x, \hat{1}]$ where $x = B_0/\gamma_{B_1}/\cdots/\gamma_{B_j}$. Index the nonzero G-blocks of x so that for some $t \in \{0, \ldots, j\}$

- $1 \leq p \leq t$ implies B_p is a nonsingleton,
- $p > t$ implies $B_p = \{a_p\}$,
- $1 \leq p < m \leq t$ implies $\max B_p < \max B_m$,
- $t < p < m$ implies $a_p < a_m$.

We now describe an increasing maximal chain c of $[x, \hat{1}]$. We consider the cases $x \neq \hat{0}$ and $x = \hat{0}$. In each case, we will prove that c is increasing, that it is the only increasing chain of $[x, \hat{1}]$, and that it precedes every other maximal chain in $[x, \hat{1}]$.

Case I ($x \neq \hat{0}$). At least one of B_0 and B_1 is a nonsingleton, so we can form the chain $c : x = x_0 \lessdot \cdots \lessdot x_j = \hat{1}$ for which $x_p = (B_0 \cup B_1 \cup \cdots \cup B_p)/\gamma_{B_{p+1}}/\cdots/\gamma_{B_j}$. The label sequence of c is $(1, \max B_1) < \cdots < (1, \max B_t) < (2, a_{t+1}) < \cdots < (2, a_j)$, so c is increasing. If $|B_0| = 1$, then we are using the fact that $h < k$.

To see that c is the only increasing chain in $[x, \hat{1}]$, let $c' : x = x'_0 \lessdot \cdots \lessdot x'_d = \hat{1}$ be a different maximal chain of $[x, \hat{1}]$. We show that merges of types h, NS-NS, k, and S-NS create (possibly weak) descents in the label sequence of c'. It suffices to show that there are two covers C_1 and C_2 in c' such that C_1 precedes[3] C_2 but $\lambda(C_1) \geq \lambda(C_2)$.

- Suppose c' contains an h-merge having label $(2, \cdot)$. Then γ_{B_1} is a nonsingleton. Eventually a G-block containing[4] γ_{B_1} will be merged with the zero G-block giving label $(1, \cdot)$.
- Suppose c' contains an NS-NS merge and that the last such merge forms a block γ_B. This merge has label $(1, \max B)$. Eventually a G-block containing γ_B will be merged with the zero G-block. If this G-block is γ_B, then the label $(1, \max B)$ is repeated. If it strictly contains γ_B, then a singleton G-block is

[2] Every cover in the product corresponds to a cover in one of the factors. To say that λ_S restricts to λ_{B_0} means that the λ_{B_0}-label of the cover in $Q_{B_0}^{h,k}(G)$ is the same as the λ_S-label of the naturally corresponding cover in $[\hat{0}, y]$.

[3] We say that C_1 *precedes* C_2 in c' if C_2 is closer to $\hat{1}$ than C_1 in c'.

[4] We say that γ_B *contains* $\delta_{B'}$ and write $\delta_{B'} \subseteq \gamma_B$ if $B' \subseteq B$ and $\overline{\delta_{B'}} = \overline{\gamma_{B'}}$.

merged with γ_B giving a label $(2, \cdot)$ prior to the merge with the zero G-block which has label $(1, \cdot)$.

- Suppose c' contains a k-merge with label $(2, \cdot)$. Eventually a G-block containing the G-block that results from this merge will be merged with the zero G-block giving the label $(1, \cdot)$.
- Suppose c' contains an S-NS merge with label $(2, \cdot)$. Eventually a G-block containing the G-block formed in this merge will be merged with the zero G-block giving label $(1, \cdot)$.

In every case, a (possibly weak) descent occurs in the label sequence of c'. We conclude that every merge in an increasing chain must be of type Z-S or of type Z-NS. All Z-NS merges have label $(1, \cdot)$ and so must precede all Z-S merges which have label $(2, \cdot)$. These merges must occur in the same order as in c because of the second coordinate of the label. Therefore, c is unique.

Next we show that c is lexicographically first among the maximal chains of $[x, \hat{1}]$. It suffices to show that if $x_p \lessdot x_{p+1}$ is any cover in c and $x_p \lessdot y \neq x_{p+1}$, then $\lambda_S(x_p \lessdot x_{p+1}) < \lambda_S(x_p \lessdot y)$. We consider the cases in which the nonzero G-block in $x_p \lessdot x_{p+1}$ is a nonsingleton or a singleton.

Case I(a) ($x_p \lessdot x_{p+1}$ is a *Z-NS* merge). If $x_p \lessdot y$ is a Z-S merge, a k-merge, an h-merge, or an S-NS merge, then $\lambda_S(x_p \lessdot y) = (2, \cdot) > (1, \max B_{p+1}) = \lambda_S(x_p \lessdot x_{p+1})$ and we are done.

If $x_p \lessdot y$ is a Z-NS merge or an NS-NS merge, then let γ_{B_m} be the nonzero G-block in the merge with the largest maximum. We have $m > p+1$ so $\lambda_S(x_p \lessdot y) = (1, \max B_m) > (1, \max B_{p+1}) = \lambda_S(x_p \lessdot x_{p+1})$.

Case I(b) ($x_p \lessdot x_{p+1}$ is a *Z-S* merge). In this case, x_p consists of the zero G-block and a number of singleton G-blocks, so $x_p \lessdot y$ either is a k-merge or is a different Z-S merge. In either case, if $\gamma_{\{a_m\}}$ is the singleton in $x_p \lessdot y$ such that m is as large as possible, then $\lambda_S(x_p \lessdot y) = (2, a_m) > (2, a_{p+1}) = \lambda_S(x_p \lessdot x_{p+1})$. This concludes Case I.

Case II ($x = \hat{0}$). We form the chain $c : \hat{0} = x_0 \lessdot \cdots \lessdot x_{n-h+1} = \hat{1}$ for which

$$x_p = \{0, a_1, \ldots, a_{p+h-1}\}/\gamma_{\{a_{p+h}\}}/\cdots/\gamma_{\{a_n\}}$$

for $p = 1, \ldots, n-h+1$. Observe that the label sequence $(2, a_h) < (2, a_{h+1}) < \cdots < (2, a_n)$ of c is increasing.

Now we show that c is the only increasing chain in $[\hat{0}, \hat{1}]$. Let $c' : \hat{0} = x'_0 \lessdot \cdots \lessdot x'_d = \hat{1}$ be a maximal chain of $[\hat{0}, \hat{1}]$. If there is a k-merge in c' giving the G-block γ_B, then it has label $(2, \cdot)$. Eventually a G-block containing γ_B will be merged with the zero G-block, giving the label $(1, \cdot)$ and a descent in c'.

Thus an h-merge is the first merge in any increasing chain and the rest are Z-S merges. The first merge must involve the first h nonzero singletons, because if one of them is omitted, it will eventually be merged with the zero G-block giving a descent. The remaining singletons must be merged into the zero block in the same order as in c because of the second coordinate, so c is unique.

Finally, we show that c is lexicographically first among the maximal chains of $[\hat{0}, \hat{1}]$. The same sufficiency condition that we used in Case I applies here. Suppose $p = 0$. Then $\hat{0} \lessdot y$ is either a k-merge or an h-merge different from the one that gives x_1. In either case, $\lambda_S(\hat{0} \lessdot y) > \lambda_S(\hat{0} \lessdot x_1)$ as desired.

If $p > 0$, then x_p is of the form $\{0, a_1, \ldots, a_{p+h-1}\}/\gamma_{\{a_{p+h}\}}/\cdots/\gamma_{\{a_n\}}$. Therefore $x_p \lessdot y$ is either a k-merge or a Z-S merge involving a singleton G-block other than $\gamma_{\{a_{p+h}\}}$. In either case, $\lambda_S(x_p \lessdot y) > \lambda_S(x_p \lessdot x_{p+1})$ as desired. □

Corollary 3.2. *If $1 \leq h < k$ and $n \geq h$, then $\Delta(Q_n^{h,k}(G))$ has the homotopy type of a wedge of spheres.*

The next theorem gives another EL-shelling for $Q_n^{h,k}(G)$ that specializes to that of [Bj-Sa] for $\overline{\Pi}_n^{h,k}$ and gives new EL-shellings for Π_{n+1} and Π_{n+1}^k.

Theorem 3.3. *If $1 \leq h < k$ and $n \geq h$, then the labelling μ of $Q_n^{h,k}(G)$ defined by*

$$\mu(x \lessdot y) = \begin{cases} (1, \max B), & B_0/\gamma_B \lessdot B_0 \cup B, \\ (2, \max_i a_i), & \{0\}/\gamma_{\{a_1\}}/\cdots/\gamma_{\{a_h\}} \lessdot \{0, a_1, \ldots, a_h\}, \\ (2, a), & B_0/\gamma_{\{a\}} \lessdot B_0 \cup \{a\}, \\ (3, \max(B \cup B')), & \gamma_B/\gamma_{B'} \lessdot \gamma_{B \cup B'}, \\ (4, \max_i a_i), & \gamma_{\{a_1\}}/\cdots/\gamma_{\{a_k\}} \lessdot \gamma_{\{a_1, \ldots, a_k\}}, \\ (4, a), & \gamma_{\{a\}}/\gamma_B \lessdot \gamma_{\{a\} \cup B} \end{cases}$$

is an EL-shelling.

Proof. This proof is nearly identical to the previous one until the point at which we need to show that c is the only increasing chain in an interval of the form $[x, \hat{1}]$, where $x \neq \hat{0}$. The proof that there can be no h-merge in any increasing chain of $[x, \hat{1}]$ is the same as before. We need to show that there can be no NS-NS merges, no k-merges, and no S-NS merges in an increasing chain of $[x, \hat{1}]$.

- Suppose c' contains an NS-NS merge with label $(3, \cdot)$. Eventually a G-block containing the G-block that results from this merge will be merged with the zero G-block, giving a label $(1, \cdot)$.
- Suppose c' contains a k-merge with label $(4, \cdot)$. Eventually, a G-block containing the G-block that results from this merge will be merged with the zero G-block, giving a label $(1, \cdot)$.
- Suppose c' contains an S-NS merge with label $(4, \cdot)$. Eventually, a G-block containing the G-block that results from this merge will be merged with the zero G-block giving a label $(1, \cdot)$.

As before, we conclude that c is unique.

In showing that c is lexicographically first among the maximal chains of $[x, \hat{1}]$ we again consider the merge $x_p \lessdot x_{p+1}$ in c and any other merge $x_p \lessdot y$. Suppose first that $x_p \lessdot x_{p+1}$ is a Z-NS merge. The proof is the same as before if $x_p \lessdot y$ is another Z-NS merge, an h-merge, or a Z-S merge.

If $x_p \lessdot y$ is a k-merge or an S-NS merge, then $\mu_S(x_p \lessdot x_{p+1}) = (1, \max B_{p+1}) < (4, \cdot) = \mu_S(x_p \lessdot y)$. If $x \lessdot y$ is an NS-NS merge, then $\mu_S(x_p \lessdot x_{p+1}) = (1, \max B_{p+1}) < (3, \cdot) = \mu_S(x_p \lessdot y)$.

If $x_p \lessdot x_{p+1}$ is a Z-S merge, then as before $x_p \lessdot y$ must be either a Z-S merge involving a singleton $\gamma_{\{a_m\}}$ with $m > p+1$ or a k-merge. In the first case, $\mu_S(x_p \lessdot y) = (2, a_m) > (2, a_{p+1}) = \mu_S(x_p \lessdot x_{p+1})$. In the second case, $\mu_S(x_p \lessdot y) = (4, \cdot) > (2, a_{p+1}) = \mu_S(x_p \lessdot x_{p+1})$ as desired.

Now consider the interval $[\hat{0}, \hat{1}]$ and let c be as in the previous proof. The label sequences $\lambda_S(c)$ and $\mu_S(c)$ are identical, so c is increasing in the labelling μ_S.

Suppose a different maximal chain $c' : \hat{0} = x'_0 \lessdot \cdots \lessdot x'_d = \hat{1}$ of $[\hat{0}, \hat{1}]$ contains a type k-merge with label $(3, \cdot)$. Eventually, a G-block containing the G-block that results from this merge will be merged with the zero G-block, giving the label $(1, \cdot)$. Therefore c' cannot be increasing. The same reasoning as before shows that c' cannot be increasing when c' consists of an h-merge and some Z-S merges. We conclude that c is the unique increasing chain of $[\hat{0},\hat{1}]$.

To see that c is lexicographically first among the maximal chains of $[\hat{0},\hat{1}]$ suppose $p=0$ and $\hat{0}\lessdot y$ is a k-merge. Then $\mu_S(\hat{0}\lessdot y)=(4,\cdot)>(2,a_h)=\mu_S(\hat{0}\lessdot x_1)$. The case when $\hat{0}\lessdot y$ is an h-merge is the same as before.

If $p>0$ and $x_p\lessdot y$ is a k-merge, then $\mu_S(x_p\lessdot y)=(4,\cdot)>(2,a_{p+h})=\mu_S(x_p\lessdot x_{p+1})$. The case when $p>0$ and $x_p\lessdot y$ is a Z-S merge is the same as before. □

4. Betti numbers for $Q_n^{h,k}(G)$. We now find the rank of $\tilde{H}^l(Q_n^{h,k}(G))$, where $1\le h<k$ and $n\ge h$. This number is known as the lth reduced Betti number of $Q_n^{h,k}(G)$ and will be denoted by $\tilde{\beta}^l_{n,h,k,G}$. When $k=2$ we obtain the usual Dowling lattice whose unique nonzero Betti number $\tilde{\beta}^{n-2}_{n,1,2,G}$ is known to be $\prod_{i=0}^{n-1}(i|G|+1)$, so assume $k>2$. Theorem 1.1 and Theorem 3.1 tell us that $\tilde{\beta}^l_{n,h,k,G}$ is the number of λ-decreasing maximal chains of length $l+2$ in $Q_n^{h,k}(G)$, so we want to count these chains. Another approach would be to count the h,k-caterpillars of section 5, which are in bijection with the decreasing chains of $Q_n^{h,k}(G)$.

For a maximal chain c of $Q_n^{h,k}(G)$ let t_c denote the number of k-merges in c. Maximal chains in $Q_n^{h,k}(G)$ are of two types: those that have an h-merge and those that do not. If c has no h-merge, then t_c satisfies $1\le t_c\le\lfloor n/k\rfloor$ and the length of c is given by $l(c)=n-t_c(k-2)$. If c has one h-merge, then t_c satisfies $0\le t_c\le\lfloor(n-h)/k\rfloor$ and the length of c is given by $l(c)=n-t_c(k-2)-(h-1)$. Let $D^l_{n,h,k,G}$ and $\hat{D}^l_{n,h,k,G}$ be the number of λ-decreasing chains in $Q_n^{h,k}(G)$ of length l that have no h-merge and one h-merge, respectively.

In any λ-decreasing maximal chain c in $Q_n^{h,k}(G)$, all h-merges, k-merges, Z-S merges, and S-NS merges must come before all Z-NS merges and NS-NS merges because of the first coordinate of the labels. Following [Su-Wa] we refer to the least element of c that has no nonzero singleton G-blocks as the *pivot* of the chain. The pivot divides the chain into a *lower portion* and an *upper portion.*

THEOREM 4.1. *Suppose $1\le h<k$ and $k>2$ and $n\ge h$. If $l=n-t(k-2)$ for some t where $1\le t\le\lfloor n/k\rfloor$, then*

$$D^l_{n,h,k,G}=|G|^{n-t}\sum_{0=i_0\le\cdots\le i_t=n-tk}\left(\prod_{j=0}^{t-1}\binom{n-jk-i_j-1}{k-1}(1+j|G|)(j+1)^{i_{j+1}-i_j}\right). \tag{4.1}$$

If $l=n-t(k-2)-(h-1)$ for some t where $0\le t\le\lfloor(n-h)/k\rfloor$, then

$$\hat{D}^l_{n,h,k,G}=\begin{cases}\dbinom{n-1}{h-1} & \textit{if } t=0,\\ \displaystyle\sum_{m=kt}^{n-h}\binom{n}{m}\binom{n-m-1}{h-1}D^{m-t(k-2)}_{m,h,k,G} & \textit{if } t\ge 1.\end{cases} \tag{4.2}$$

All other values of $D^l_{n,h,k,G}$ and $\hat{D}^l_{n,h,k,G}$ are zero.

Proof. We will count λ-decreasing chains c for which $t_c=t$, i.e., chains having pivots with t nonzero G-blocks. First we count λ-decreasing chains having no h-merge. In such a chain, the zero G-block cannot participate in any merge in the lower portion. The k-merges of c divide the lower portion of c into t segments. For $0\le j\le t$ let i_j be the number of merges of a singleton with one other G-block that occur while there are j or fewer nonzero nonsingleton G-blocks, so that $0=i_0\le i_1\le\cdots\le i_t=n-tk$.

When performing the $(j+1)$st k-merge we must use the largest remaining singleton together with $k-1$ of the other $n-jk-i_j-1$ available singletons. $|G|^{k-1}$ different G-blocks can be formed from a fixed set of k nonzero singleton G-blocks, so the number of choices for the $(j+1)$st k-merge is

$$|G|^{k-1}\binom{n-jk-i_j-1}{k-1}.$$

There are $|G|(j+1)$ ways to perform each singleton merge when $j+1$ nonsingleton G-blocks are present, so there are a total of $[|G|(j+1)]^{i_{j+1}-i_j}$ ways to merge the $i_{j+1}-i_j$ singletons that will be merged between the $(j+1)$st and the $(j+2)$th k-merge. The total number of choices for the lower portion of c is therefore

$$\begin{aligned} &\sum_{0=i_0\leq\cdots\leq i_t=n-kt}\left(\prod_{j=0}^{t-1}|G|^{k-1}\binom{n-jk-i_j-1}{k-1}[|G|(j+1)]^{i_{j+1}-i_j}\right)\\ &= |G|^{n-t}\sum_{0=i_0\leq\cdots\leq i_t=n-tk}\left(\prod_{j=0}^{t-1}\binom{n-jk-i_j-1}{k-1}(j+1)^{i_{j+1}-i_j}\right). \end{aligned} \tag{4.3}$$

After the pivot there comes a sequence of t merges of types Z-NS and NS-NS. The nonzero G-block with the largest maximum must be involved in each of these merges to ensure that c is λ-decreasing. For the jth merge above the pivot, there is one way to merge this G-block with the zero G-block and $|G|$ ways to merge it with each of the $t-j$ other nonzero G-blocks. Thus there are a total of $1+|G|(t-j)$ choices for the jth merge and a total of

$$\prod_{j=1}^{t-1}(1+j|G|) \tag{4.4}$$

choices for the upper portion of the chain. Taking the product of (4.3) with (4.4) we obtain equation (4.1).

To obtain the formula for $\hat{D}^l_{n,h,k}$ when $t=0$, observe that the h-merge must be the first merge and that all merges after that are between singletons and the zero G-block. The h-merge must involve the largest singleton G-block and $h-1$ other nonzero singletons. There are $\binom{n-1}{h-1}$ ways to perform the h-merge. The singletons that remain must be merged with the zero G-block in decreasing order of first coordinate, so no further choice is involved. This concludes the first part of (4.2).

Now consider the case when $t\geq 1$ and there is an h-merge. Let $T\subseteq\{1,\dots,n\}$ with $|T|=m$ where $kt\leq m\leq n-h$. Let $S=T\cup\{0\}$. We count over all such sets S the number of pairs of decreasing chains (c_1,c_2) for which c_1 is in $Q^{h,k}_S(G)$ and has no h-merge and t k-merges and for which c_2 is in $Q^{h,k}_{\{0,\dots,n\}\setminus T}(G)$ and consists of one h-merge and no k-merges. Such pairs are in bijection with decreasing chains of $Q^{h,k}_n(G)$ having t k-merges and one h-merge. Merges in c_1 correspond to k-merges, S-NS merges, NS-NS merges, and Z-NS merges in c. Merges in c_2 correspond to h-merges and Z-S merges in c.

For example, let $S = \{0, 1, 2, 3, 5, 6, 9, 10, 11, 13, 15\}$. The λ-decreasing chain

$$\begin{array}{rcl}
c & = & \text{0/1/2/3/4/5/6/7/8/9/10/11/12/13/14/15} \\
 & \lessdot & \text{0/1/2/}\textbf{3 9 11 15}\text{/4/5/6/7/8/10/12/13/14} \\
 & \lessdot & \textbf{0 4 14}\text{/1/2/3 9 11 15/5/6/7/8/10/12/13} \\
 & \lessdot & \text{0 4 14/1/}\textbf{2 6 10 13}\text{/3 9 11 15/5/7/8/12} \\
 & \lessdot & \textbf{0 4 12 14}\text{/1/2 6 10 13/3 9 11 15/5/7/8} \\
 & \lessdot & \textbf{0 4 8 12 14}\text{/1/2 6 10 13/3 9 11 15/5/7} \\
 & \lessdot & \textbf{0 4 7 8 12 14}\text{/1/2 6 10 13/3 9 11 15/5} \\
 & \lessdot & \text{0 4 7 8 12 14/1/2 6 10 13/}\textbf{3 5 9 11 15} \\
 & \lessdot & \text{0 4 7 8 12 14/}\textbf{1 2 6 10 13}\text{/3 5 9 11 15} \\
 & \lessdot & \textbf{0 3 4 5 7 8 9 11 12 14 15}\text{/1 2 6 10 13} \\
 & \lessdot & \textbf{0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15}
\end{array}$$

in $Q^{2,4}_{15}((e))$ corresponds to the pair of λ-decreasing chains

$$\begin{array}{rclcl}
c_1 & = & \text{0/1/2/3/5/6/9/10/11/13/15} & \lessdot & \text{0/1/2/}\textbf{3 9 11 15}\text{/5/6/10/13} \\
 & \lessdot & \text{0/1/}\textbf{2 6 10 13}\text{/3 9 11 15/5} & \lessdot & \text{0/1/2 6 10 13/}\textbf{3 5 9 11 15} \\
 & \lessdot & \text{0/}\textbf{1 2 6 10 13}\text{/3 5 9 11 15} & \lessdot & \textbf{0 3 5 9 11 15}\text{/1 2 6 10 13} \\
 & \lessdot & \textbf{0 1 2 3 5 6 9 10 11 13 15} & &
\end{array}$$

and

$$\begin{array}{rclclcl}
c_2 & = & \text{0/4/7/8/12/14} & \lessdot & \textbf{0 4 14}\text{/7/8/12} & \lessdot & \textbf{0 4 12 14}\text{/7/8} \\
 & \lessdot & \textbf{0 4 8 12 14}\text{/7} & \lessdot & \textbf{0 4 7 8 12 14} & &
\end{array}$$

in $Q^{2,4}_{\{0,1,2,3,5,6,9,10,11,13,15\}}((e))$ and $Q^{2,4}_{\{0,4,7,8,12,14\}}((e))$, respectively. Here and elsewhere, each newly formed $\{e\}$-block is displayed in boldface for readability.

There are $\binom{n}{m}$ ways to choose $T \subseteq \{1, \ldots, n\}$ with $|T| = m$. As seen in (4.1) there are $D^{m-t(k-2)}_{m,h,k,G}$ decreasing chains in $Q^{h,k}_S(G)$ for which the number of k-merges is t and the number of h-merges is 0, so there are $\binom{n}{m} D^{m-t(k-2)}_{m,h,k,G}$ chains that can serve as c_1. As seen in the first part of (4.2) there are $\binom{n-m-1}{h-1}$ decreasing chains in $Q^{h,k}_{\{0,\ldots,n\}\setminus T}(G)$ that can serve as c_2. Thus we have

$$\hat{D}^l_{n,h,k,G} = \sum_{m=kt}^{n-h} \binom{n}{m} \binom{n-m-1}{h-1} D^{m-t(k-2)}_{m,h,k,G}$$

as desired for the second part of (4.2). □

Taken together, Theorem 1.1 and Theorem 4.1 give the following corollary. When $1 < h < k - 1$, the length l of a chain determines whether the chain includes an h-merge or not. When $h = 1$ or $h = k - 1$, we must count chains of length l that include an h-merge as well as those that do not.

Corollary 4.2. *If $1 \leq h < k$ and $n \geq h$ and $k > 2$, then $Q^{h,k}_n(G)$ has the homotopy type of a wedge of spheres. Its integral homology groups are free with* *reduced Betti numbers*

$$\tilde{\beta}^l_{n,h,k,G} = \begin{cases} D^{l+2}_{n,h,k,G} & \textit{if } 1 < h < k-1 \textit{ and } l = n - t(k-2) - 2 \\ & \textit{for some } t \textit{ where } 1 \le t \le \lfloor n/k \rfloor, \\ \hat{D}^{l+2}_{n,h,k,G} & \textit{if } 1 < h < k-1 \textit{ and } l = n - t(k-2) - (h-1) - 2 \\ & \textit{for some } t \textit{ where } 0 \le t \le \lfloor (n-h)/k \rfloor, \\ D^{l+2}_{n,h,k,G} + \hat{D}^{l+2}_{n,h,k,G} & \textit{if } h = 1 \textit{ and } l = n - t(k-2) - 2 \\ & \textit{for some } t \textit{ where } 0 \le t \le \lfloor n/k \rfloor, \\ D^{l+2}_{n,h,k,G} + \hat{D}^{l+2}_{n,h,k,G} & \textit{if } h = k-1 \textit{ and } l = n - t(k-2) - 2 \\ & \textit{for some } t \textit{ where } 1 \le t \le \lfloor (n+1)/k \rfloor, \\ 0 & \textit{otherwise.} \end{cases}$$

5. The h, k-caterpillar basis for homology. We describe dual bases for the homology and cohomology groups of $Q^{h,k}_n(G)$ in terms of trees which we call h, k-caterpillars. The cohomology basis is obtained from the λ-decreasing chains (cf. Theorem 1.1). The homology basis is modelled on the caterpillar basis for Π^k_n of [Wa2].

There are two kinds of h, k-caterpillars. The first kind corresponds to decreasing chains with no h-merge, and the second kind corresponds to decreasing chains with one h-merge. h, k-caterpillars are built up from k-caterpillars, which in turn are built up from k-stars.

A *k-star* is a tree whose vertices are subsets of $\{1, \ldots, n\} \times G$ such that

- no element of $\{1, \ldots, n\}$ appears twice as the first coordinate of an element of a vertex of the k-star;
- one vertex, called the *root*, is of cardinality $k-1$, and the remaining vertices, called *legs*, are of cardinality 1;
- there is at least one leg whose element has a larger first coordinate than every element of the root;
- each leg is of degree one and is adjacent to the root.

For example, let $n = 100$ and $G = C_2$. Figure 5.1 shows a 4-star. Its legs are $\{1\}$, $\{\overline{6}\}$, $\{\overline{15}\}$, and $\{25\}$, and its root is $\{4, \overline{7}, 10\}$.

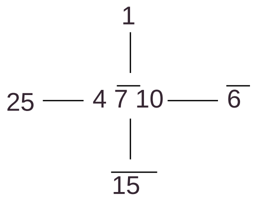


FIG. 5.1. *A 4-star.*

In our figures we adopt the notational conventions that apply to our examples, i.e., dropping curly brackets and commas from sets, putting second coordinates over first coordinates, etc.

A *k-caterpillar* is a tree such that we have the following:

- Its vertex set is the disjoint union of the vertex sets of one or more k-stars.
- No element of $\{1, \ldots, n\}$ appears twice as the first coordinate of an element of a vertex.
- Its edge set contains the edge sets of the k-stars in the union.

- The subgraph induced by the roots of the k-stars, called the *spine*, is a path. The requirement that the k-caterpillar is a tree implies that each edge is either an edge of one of the k-stars or joins the roots of two k-stars.
- The pair having the smallest first coordinate among all elements of the vertices of the k-caterpillar has a second coordinate e.
- The k-star to which the leg with the largest first coordinate belongs is called the *maximum k-star*; its root is at one end of the spine.

The legs of the k-stars are also called legs of the k-caterpillar. For example, let $n = 100$ and $G = C_2$. The tree depicted in Figure 5.2 is a 4-caterpillar. Its maximum 4-star is the rightmost star.

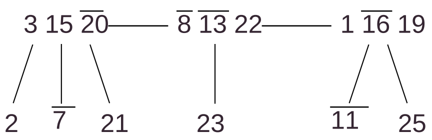


FIG. 5.2. *A 4-caterpillar.*

An *h, k-caterpillar of the first type* is a tree such that

- $v_0 = \{0\}$ is a vertex and the remaining vertices are subsets of $\{1, \dots, n\} \times G$,
- each element of $\{1, \dots, n\}$ appears exactly once as the first coordinate of an element of a vertex,
- each component of the induced subforest on the vertices other than v_0 is a k-caterpillar,
- each k-caterpillar in the subforest is joined to v_0 at the root of the k-star farthest from the maximum k-star.

The value of h cannot be inferred from an h, k-caterpillar of the first type since the corresponding chains have no h-merge.

For example, if $n = 27$ and $G = C_2$, then the tree depicted in Figure 5.3 is a $2, 4$-caterpillar of the first type.

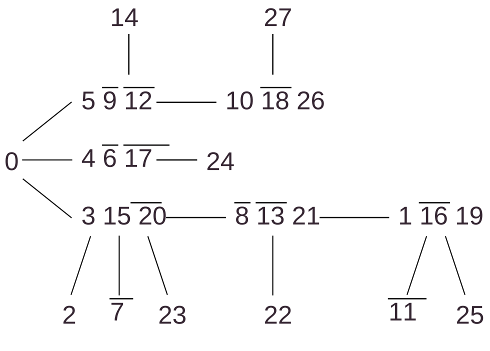


FIG. 5.3. *A $2, 4$-caterpillar of the first type.*

An *h, k-caterpillar of the second type* is a tree such that we have the following:

- $v_0 = \{0, (a_1, e), \dots, (a_{h-1}, e)\}$ is a vertex, where $a_i \in \{1, \dots, n\}$ for $1 \leq i < h$.
- Every vertex adjacent to v_0 is either a singleton from $\{1, \dots, n\} \times \{e\}$ or is the root of the k-star furthest from the maximum k-star in a k-caterpillar.

The singleton elements adjacent to v_0 are also called legs.
- There is a singleton adjacent to v_0 the first coordinate of which is larger than the first coordinate of every pair in v_0.
- Each element of $\{1, \ldots, n\}$ appears exactly once as the first coordinate of an element of a vertex.

For example, if $n = 20$ and $G = C_2$, then the tree depicted in Figure 5.4 is a $2, 4$-caterpillar of the second type.

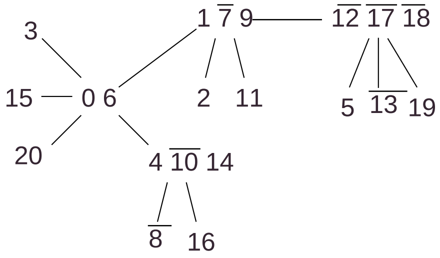


FIG. 5.4. *A* $2, 4$-*caterpillar of the second type.*

Any set S of edges of an h, k-caterpillar κ of either type gives an element π_S of $Q_n^{h,k}(G)$ whose G-blocks are obtained from the union of the vertices in the connected components of κ corresponding to S and from breaking isolated vertices of size h or $k - 1$ into singleton G-blocks.

For example, if S is the set of edges of the $2, 4$-caterpillar depicted in Figure 5.5, then the corresponding C_2-partition is

$$\pi_S = 0\ 1\ 6\ 7\ 9\ 11\ 15/2/3/4/5/\overline{8}/\overline{10}/\overline{12}\ \overline{17}\ \overline{18}\ 19/\overline{13}/14/16/20$$

or, using the canonical representative,

$$\pi_S = 0\ 1\ 6\ 7\ 9\ 11\ 15/2/3/4/5/8/10/12\ 17\ 18\ \overline{19}/13/14/16/20.$$

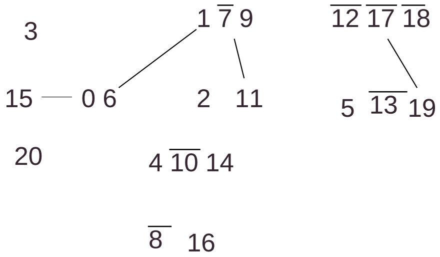


FIG. 5.5. *A* $2, 4$-*caterpillar with some edges deleted.*

Furthermore, $S_1 \subseteq S_2$ if and only if $\pi_{S_1} \leq \pi_{S_2}$. For each h, k-caterpillar κ, we therefore obtain a sublattice Q_κ of $Q_n^{h,k}(G)$ that is isomorphic to the Boolean lattice on the number of edges of κ.

Let ρ_κ be the fundamental cycle of spherical complex Q_κ; this cycle is unique up to sign. We claim that ρ_κ contains exactly one summand $\hat{c}_\kappa$ such that c_κ is λ-decreasing. c_κ can be obtained as follows. The smallest element of c_κ corresponds to no edges of κ being included. To get the next element of c_κ find the leg whose element has the largest first coordinate and add the edge incident to that leg. Continue by adding edges incident to singleton vertices in decreasing order of the first coordinate of the element of the singleton vertices. When all edges incident to legs have been added, we have reached the pivot.

Next find the k-caterpillar having the pair with the largest first coordinate as an element of one of its vertices and add the edges in the spine of that k-caterpillar, starting at the maximum k-star and working towards v_0. Then find the k-caterpillar having the pair with the next largest first coordinate as an element of one of its vertices and repeat. Continue until all edges of κ are added.

For example, if κ is the 2, 4-caterpillar in Figure 5.3, then c_κ is the chain

$$\begin{array}{l}
0/1/2/3/4/5/\overline{6}/\overline{7}/\overline{8}/\overline{9}/10/\overline{11}/\overline{12}/\overline{13}/14/15/\overline{16}/\overline{17}/\overline{18}/19/\overline{20}/21/22/23/24/25/26/27\\
\lessdot 0/1/2/3/4/5/\overline{6}/\overline{7}/\overline{8}/\overline{9}/\mathbf{10\ \overline{18}\ 26\ 27}/\overline{11}/\overline{12}/\overline{13}/14/15/\overline{16}/\overline{17}/19/\overline{20}/21/22/23/24/25\\
\lessdot 0/\mathbf{1\ \overline{16}\ 19\ 25}/2/3/4/5/\overline{6}/\overline{7}/\overline{8}/\overline{9}/10\ \overline{18}\ 26\ 27/\overline{11}/\overline{12}/\overline{13}/14/15/\overline{17}/\overline{20}/21/22/23/24\\
\lessdot 0/1\ \overline{16}\ 19\ 25/2/3/\mathbf{4\ \overline{6}\ \overline{17}\ 24}/5/\overline{7}/\overline{8}/\overline{9}/10\ \overline{18}\ 26\ 27/\overline{11}/\overline{12}/\overline{13}/14/15/\overline{20}/21/22/23\\
\lessdot 0/1\ \overline{16}\ 19\ 25/2/\mathbf{3\ 15\ \overline{20}\ 23}/4\ \overline{6}\ \overline{17}\ 24/5/\overline{7}/\overline{8}/\overline{9}/10\ \overline{18}\ 26\ 27/\overline{11}/\overline{12}/\overline{13}/14/21/22\\
\lessdot 0/1\ \overline{16}\ 19\ 25/2/3\ 15\ \overline{20}\ 23/4\ \overline{6}\ \overline{17}\ 24/5/\overline{7}/\mathbf{\overline{8}\ \overline{13}\ 21\ 22}/\overline{9}/10\ \overline{18}\ 26\ 27/\overline{11}/\overline{12}/14\\
\lessdot 0/1\ \overline{16}\ 19\ 25/2/3\ 15\ \overline{20}\ 23/4\ \overline{6}\ \overline{17}\ 24/\mathbf{5\ \overline{9}\ \overline{12}\ 14}/\overline{7}/\overline{8}\ \overline{13}\ 21\ 22/10\ \overline{18}\ 26\ 27/\overline{11}\\
\lessdot 0/\mathbf{1\ \overline{11}\ \overline{16}\ 19\ 25}/2/3\ 15\ \overline{20}\ 23/4\ \overline{6}\ \overline{17}\ 24/5\ \overline{9}\ \overline{12}\ 14/\overline{7}/\overline{8}\ \overline{13}\ 21\ 22/10\ \overline{18}\ 26\ 27\\
\lessdot 0/1\ \overline{11}\ \overline{16}\ 19\ 25/2/\mathbf{3\ \overline{7}\ 15\ \overline{20}\ 23}/4\ \overline{6}\ \overline{17}\ 24/5\ \overline{9}\ \overline{12}\ 14/\overline{8}\ \overline{13}\ 21\ 22/10\ \overline{18}\ 26\ 27\\
\lessdot 0/1\ \overline{11}\ \overline{16}\ 19\ 25/\mathbf{2\ 3\ \overline{7}\ 15\ \overline{20}\ 23}/4\ \overline{6}\ \overline{17}\ 24/5\ \overline{9}\ \overline{12}\ 14/\overline{8}\ \overline{13}\ 21\ 22/10\ \overline{18}\ 26\ 27\\
\lessdot 0/1\ \overline{11}\ \overline{16}\ 19\ 25/2\ 3\ \overline{7}\ 15\ \overline{20}\ 23/4\ \overline{6}\ \overline{17}\ 24/\mathbf{5\ \overline{9}\ 10\ \overline{12}\ 14\ \overline{18}\ 26\ 27}/\overline{8}\ \overline{13}\ 21\ 22\\
\lessdot \mathbf{0\ 5\ 9\ 10\ 12\ 14\ 18\ 26\ 27}/1\ \overline{11}\ \overline{16}\ 19\ 25/2\ 3\ \overline{7}\ 15\ \overline{20}\ 23/4\ \overline{6}\ \overline{17}\ 24/\overline{8}\ \overline{13}\ 21\ 22\\
\lessdot 0\ 5\ 9\ 10\ 12\ 14\ 18\ 26\ 27/\mathbf{1\ \overline{8}\ \overline{11}\ \overline{13}\ \overline{16}\ 19\ 21\ 22\ 25}/2\ 3\ \overline{7}\ 15\ \overline{20}\ 23/4\ \overline{6}\ \overline{17}\ 24\\
\lessdot 0\ 5\ 9\ 10\ 12\ 14\ 18\ 26\ 27/\mathbf{1\ 2\ 3\ \overline{7}\ \overline{8}\ \overline{11}\ \overline{13}\ 15\ \overline{16}\ 19\ \overline{20}\ 21\ 22\ 23\ 25}/4\ \overline{6}\ \overline{17}\ 24\\
\lessdot \mathbf{0\ 1\ 2\ 3\ 5\ 7\ 8\ 9\ 10\ 11\ 12\ 13\ 14\ 15\ 16\ 18\ 19\ 20\ 21\ 22\ 23\ 25\ 26\ 27}/4\ \overline{6}\ \overline{17}\ 24\\
\lessdot \mathbf{0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ 13\ 14\ 15\ 16\ 17\ 18\ 19\ 20\ 21\ 22\ 23\ 24\ 25\ 26\ 27},
\end{array}$$

and if C is the 2, 4-caterpillar of the second type in Figure 5.4, then c_κ is the chain

$$\begin{array}{l}
0/1/2/3/4/5/\overline{6}/\overline{7}/\overline{8}/9/\overline{10}/11/\overline{12}/\overline{13}/14/15/16/\overline{17}/\overline{18}/19/\overline{20}\\
\lessdot \mathbf{0\ 6\ 20}/1/2/3/4/5/\overline{7}/\overline{8}/9/\overline{10}/11/\overline{12}/\overline{13}/14/15/16/\overline{17}/\overline{18}/19\\
\lessdot 0\ 6\ 20/1/2/3/4/5/\overline{7}/\overline{8}/9/\overline{10}/11/\mathbf{\overline{12}\ \overline{17}\ \overline{18}\ 19}/\overline{13}/14/15/16\\
\lessdot 0\ 6\ 20/1/2/3/\mathbf{4\ \overline{10}\ 14\ 16}/5/\overline{7}/\overline{8}/9/11/\overline{12}\ \overline{17}\ \overline{18}\ 19/\overline{13}/15\\
\lessdot \mathbf{0\ 6\ 15\ 20}/1/2/3/4\ \overline{10}\ 14\ 16/5/\overline{7}/\overline{8}/9/11/\overline{12}\ \overline{17}\ \overline{18}\ 19/\overline{13}\\
\lessdot 0\ 6\ 15\ 20/1/2/3/4\ \overline{10}\ 14\ 16/5/\overline{7}/\overline{8}/9/11/\mathbf{\overline{12}\ \overline{13}\ \overline{17}\ \overline{18}\ 19}\\
\lessdot 0\ 6\ 15\ 20/\mathbf{1\ \overline{7}\ 9\ 11}/2/3/4\ \overline{10}\ 14\ 16/5/\overline{8}/\overline{12}\ \overline{13}\ \overline{17}\ \overline{18}\ 19\\
\lessdot 0\ 6\ 15\ 20/1\ \overline{7}\ 9\ 11/2/3/\mathbf{4\ \overline{8}\ \overline{10}\ 14\ 16}/5/\overline{12}\ \overline{13}\ \overline{17}\ \overline{18}\ 19\\
\lessdot 0\ 6\ 15\ 20/1\ \overline{7}\ 9\ 11/2/3/4\ \overline{8}\ \overline{10}\ 14\ 16/\mathbf{5\ \overline{12}\ \overline{13}\ \overline{17}\ \overline{18}\ 19}\\
\lessdot \mathbf{0\ 3\ 6\ 15\ 20}/1\ \overline{7}\ 9\ 11/2/4\ \overline{8}\ \overline{10}\ 14\ 16/5\ \overline{12}\ \overline{13}\ \overline{17}\ \overline{18}\ 19\\
\lessdot 0\ 3\ 6\ 15\ 20/\mathbf{1\ 2\ \overline{7}\ 9\ 11}/4\ \overline{8}\ \overline{10}\ 14\ 16/5\ \overline{12}\ \overline{13}\ \overline{17}\ \overline{18}\ 19\\
\lessdot 0\ 3\ 6\ 15\ 20/\mathbf{1\ 2\ 5\ \overline{7}\ 9\ 11\ \overline{12}\ \overline{13}\ \overline{17}\ \overline{18}\ 19}/4\ \overline{8}\ \overline{10}\ 14\ 16\\
\lessdot \mathbf{0\ 1\ 2\ 3\ 5\ 6\ 7\ 9\ 11\ 12\ 13\ 15\ 17\ 18\ 19\ 20}/4\ \overline{8}\ \overline{10}\ 14\ 16\\
\lessdot \mathbf{0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ 13\ 14\ 15\ 16\ 17\ 18\ 19\ 20}.
\end{array}$$

The correspondence between h, k-caterpillars and λ-decreasing chains also goes the other way. That is, to each λ-decreasing chain $c : \hat{0} = x_0 \lessdot \cdots \lessdot x_d = \hat{1}$ of $Q_n^{h,k}(G)$ there corresponds an h, k-caterpillar κ_c which is obtained by the following construction. Each merge in c corresponds to an edge in κ_c, and the merges determine the first coordinates of the vertices of κ_c as described below. The second coordinate is determined by the canonical representative of the smaller G-partition in the Z-NS merge involving a pair with the same first coordinate. For example, if c contains the merge $0/1\overline{2}3\overline{4} \lessdot 01234$, then 1, $\overline{2}$, 3, and $\overline{4}$ are elements of vertices of κ_c. In the case of a Z-S merge or an h-merge, the second coordinate is e.

- If c contains a k-merge $\gamma_{\{a_1\}}/\cdots/\gamma_{\{a_k\}} \lessdot \gamma_{\{a_1,\ldots,a_k\}}$ where $a_1 < \cdots < a_k$, then $\gamma_{\{a_1,\ldots,a_{k-1}\}}$ is a root of κ_c and this root has a leg $\gamma_{\{a_k\}}$.
- If c contains an h-merge $\{0\}/\gamma_{\{a_1\}}/\cdots/\gamma_{\{a_h\}} \lessdot \{0, a_1, \ldots, a_h\}$ with $a_1 < \cdots < a_h$, then $v_0 = \{0, (a_1, e), \ldots, (a_{h-1}, e)\}$ and the singleton $\{(a_h, e)\}$ is incident to v_0 in κ_c. Otherwise, $v_0 = \{0\}$.
- Suppose c contains an S-NS merge $\gamma_{\{a\}}/\gamma_B \lessdot \gamma_{\{a\}\cup B}$. Among the earlier merges involving the elements of γ_B there is exactly one k-merge. Let $\gamma_{\{a\}}$ be a leg incident to the root corresponding to this k-merge in κ_c.
- If c contains a Z-S merge $B_0/\gamma_{\{a\}} \lessdot B_0 \cup \{a\}$, then $\{(a, e)\}$ is incident to v_0 in κ_c. Note that a Z-S merge must be preceded by an h-merge.
- Suppose c contains an NS-NS merge $\gamma_B/\gamma_{B'} \lessdot \gamma_{B\cup B'}$. Among the earlier merges involving elements of one of γ_B and $\gamma_{B'}$ there is exactly one k-merge. Suppose this is true of γ_B. There may be several k-merges among the earlier merges involving elements of $\gamma_{B'}$, and these k-merges correspond to the roots of a sub-k-caterpillar[5] of κ_c containing the maximum k-star. Join the root corresponding to the k-merge involving elements of γ_B with the root of the k-caterpillar corresponding to $\gamma_{B'}$ that is farthest from the maximum k-star of that k-caterpillar.
- Suppose c contains a Z-NS merge $B_0/\gamma_B \lessdot B_0 \cup B$. Then an edge joins v_0 to the root of κ_c farthest from the maximum k-star of the k-caterpillar of κ_c whose vertices contain the elements of γ_B.

The reader may check that $\kappa_{c_\kappa} = \kappa$ for each λ-decreasing chain c. The two previous examples viewed in "reverse" serve as examples of this construction.

Let $d_\kappa = 0$ if v_0 is a singleton, and let $d_\kappa = 1$ if the cardinality of v_0 is greater than 1. Let t_κ be the number of roots of κ. The next theorem follows from the above discussion and Proposition 1.1 of [Wa1].

THEOREM 5.1. *Let $1 \leq h < k$ and $n \geq h$. Then*

$$\{\rho_\kappa \mid \kappa \textit{ is an } h,k\textit{-caterpillar}\} \quad \textit{and} \quad \{c_\kappa \mid \kappa \textit{ is an } h,k\textit{-caterpillar}\}$$

are dual bases for $\tilde{H}_l(Q_n^{h,k}(G))$ and $\tilde{H}^l(Q_n^{h,k}(G))$, respectively. The κ's are constrained to satisfy $l = n - t_\kappa(k-2) - d_\kappa(h-1) - 2$.

6. The subspace arrangement $\mathcal{B}_{n,m}^{h,k}$ and the cohomology of its complement.

A *subspace arrangement* of a vector space V is a set $\mathcal{A} = \{W_1, \ldots, W_r\}$ of subspaces of V. In this paper, we will be working with subspace arrangements of $\mathbf{C}^n$, which will be viewed as $\mathbf{R}^{2n}$ when necessary. The *intersection lattice* of $\mathcal{A}$, denoted $L(\mathcal{A})$, is defined to be the set of intersections of elements of $\mathcal{A}$ ordered by reverse inclusion.

Let ω be a primitive mth root of unity. The hyperplane arrangement $\mathcal{A}_{n,m}$ is defined in [Or-So] to be the set of hyperplanes of the form $\{\mathbf{z} \in \mathbf{C}^n \mid z_p = 0\}$, where $1 \leq p \leq n$, together with the hyperplanes of the form $\{\mathbf{z} \in \mathbf{C}^n \mid z_p - \omega^r z_q = 0\}$, where $1 \leq p < q \leq n$ and $0 \leq r < m$.

$\mathcal{A}_{n,1}$ is the complexified braid arrangement, also known as the Coxeter arrangement of type A_n, whose intersection lattice is the partition lattice Π_{n+1}. The ar-

[5] A sub-k-caterpillar κ of a k-caterpillar κ' is a k-caterpillar that is an induced subgraph of κ'.

rangement $\mathcal{A}_{n,2}$ is the Coxeter arrangement of type B_n whose intersection lattice is the signed partition lattice $\overline{\Pi}_n$. The Dowling lattice $Q_n(C_m)$ is the lattice of intersections $L(\mathcal{A}_{n,m})$ of the arrangement $\mathcal{A}_{n,m}$.

Define the h, k-equal Dowling subspace arrangement $\mathcal{B}^{h,k}_{n,m}$ to be the set of subspaces of the form $\{\mathbf{z} \in \mathbf{C}^n \mid z_{j_1} = \cdots = z_{j_h} = 0\}$, where $1 \leq j_1 < \cdots < j_h \leq n$, together with subspaces of the form $\{\mathbf{z} \in \mathbf{C}^n \mid \omega^{i_1} z_{j_1} = \cdots = \omega^{i_k} z_{j_k}\}$, where $0 \leq i_l < m$ for $l = 1, \dots, k$ and $1 \leq j_1 < \cdots < j_k \leq n$. Observe that $\mathcal{B}^{1,2}_{n,m} = \mathcal{A}_{n,m}$, that $\mathcal{B}^{k-1,k}_{n,1}$ is the type A k-equal subspace arrangement, and that $\mathcal{B}^{h,k}_{n,2}$ is the type B h, k-equal subspace arrangement.

Proposition 6.1. *If $1 \leq h < k$ and $n \geq h$, then $L(\mathcal{B}^{h,k}_{n,m}) = Q^{h,k}_n(C_m)$.*

Let $M^{h,k}_{n,m} = \mathbf{C}^n \setminus \cup_{A \in \mathcal{B}^{h,k}_{n,m}} A$. We will use our knowledge of the topology of $Q^{h,k}_n(C_m) = L(\mathcal{B}^{h,k}_{n,m})$ to describe the cohomology of $M^{h,k}_{n,m}$. The following theorem, known as the Goresky–MacPherson formula, shows that it suffices to understand the homology of the posets $[\hat{0}, y]$ for each $y \in Q^{h,k}_n(C_m)$ in order to understand the cohomology of $M^{h,k}_{n,m}$. For an element y of a poset P we denote the dth reduced homology of the subposet $[\hat{0}, y]$ of P by $\tilde{H}_d(\hat{0}, y)$.

Theorem 6.2 (see [Go-Ma]). *Let $\mathcal{A}$ be a subspace arrangement in $\mathbf{R}^n$ with intersection lattice L. Then for all dimensions d,*

$$\tilde{H}^d\left(\mathbf{R}^n \setminus \bigcup_{A \in \mathcal{A}} A\right) = \bigoplus_{y \in L \setminus \hat{0}} \tilde{H}_{\operatorname{codim}(y)-d-2}(\hat{0}, y).$$

In order to apply this result, we will view $A \in \mathcal{B}^{h,k}_{n,m}$ as a subset of $\mathbf{R}^{2n}$ and $M^{h,k}_{n,m}$ as $\mathbf{R}^{2n} \setminus \cup_{A \in \mathcal{B}^{h,k}_{n,m}} A$. In this context, codim$(y)$ refers to the real codimension of y as a subset of R^{2n}.

Since the intervals $[\hat{0}, y]$ are isomorphic to the product of an h, k-equal Dowling lattice with a product of k-equal partition lattices, the following results will be useful. Let $\tilde{\beta}^d_{n,k}$ denote the d-dimensional reduced Betti number of Π^k_n.

Theorem 6.3 (see [Bj-We]). *If $2 < k \leq n$, then Π^k_n has the homotopy type of a wedge of spheres, so its homology groups are free. Furthermore, $\tilde{\beta}^d_{n,k} \neq 0$ if and only if $d = n - 3 - t(k-2)$ for some t with $1 \leq t \leq \lfloor n/k \rfloor$, and in that case*

$$\tilde{\beta}^d_{n,k} = (t-1)! \sum_{0=i_0 \leq \cdots \leq i_t = n-tk} \prod_{j=0}^{t-1} \binom{n - jk - i_j - 1}{k-1} (j+1)^{i_{j+1}-i_j}.$$

Let $\tilde{\beta}^d(P)$ denote the d-dimensional reduced Betti number of the poset P. The next result follows from the Künneth formula and results of Quillen [Qu] and Walker [Wk].

Proposition 6.4. *Let P_i be a bounded finite poset for $i = 1, \dots, j$. Then*

$$\tilde{\beta}^d(P_1 \times \cdots \times P_j) = \sum_{r_1 + \cdots + r_j = d - 2(j-1)} \tilde{\beta}^{r_1}(P_1) \cdots \tilde{\beta}^{r_j}(P_j).$$

Recall that the d-dimensional reduced Betti numbers $\tilde{\beta}^d_{n,m,h,k}$ of $Q^{h,k}_n(C_m)$ were computed in Corollary 4.2. Define

$$\tilde{\beta}^d_{0,m,h,k} = \begin{cases} 1 & \text{if } d = -2, \\ 0 & \text{else.} \end{cases} \tag{6.1}$$

Theorem 6.5. *Suppose* $1 \leq h < k$ *and* $k > 2$ *and* $n \geq h$. *Let* $y = B_0/\gamma_{B_1}/\cdots/\gamma_{B_j} \in Q_n^{h,k}(C_m)$ *be such that* $\gamma_{B_1},\ldots,\gamma_{B_p}$ *are all of the nonzero nonsingleton* C_m*-blocks of* y. *Set* $a_i = |B_i|$ *for* $i = 0,1,\ldots,p$. *Then we have the following:*

1. $\tilde{\beta}^d(\hat{0},y) = \sum_{r_0+r_1+\cdots+r_p=d-2p} \tilde{\beta}^{r_0}_{a_0-1,m,h,k}\tilde{\beta}^{r_1}_{a_1,k}\cdots\tilde{\beta}^{r_p}_{a_p,k}$. *Here the* r_i*'s may assume any integer values.*
2. $\tilde{\beta}^d(\hat{0},y)$ *is nonzero if and only if one of the following hold:*
   (a) $a_0 = 1$ *and* $d = \sum_{i=1}^p a_i - p - 2 - t(k-2)$ *for some* t *where* $p \leq t \leq \sum_{i=1}^p \lfloor a_i/k \rfloor$.
   (b) $a_0 > 1$ *and* $1 < h < k-1$ *and* $d = \sum_{i=0}^p a_i - p - 3 - t(k-2)$ *for some* t *where* $p+1 \leq t \leq \lfloor (a_0-1)/k \rfloor + \sum_{i=1}^p \lfloor a_i/k \rfloor$.
   (c) $a_0 > 1$ *and* $1 < h < k-1$ *and* $d = \sum_{i=0}^p a_i - p - 3 - t(k-2) - (h-1)$ *for some* t *where* $p \leq t \leq \lfloor (a_0-h-1)/k \rfloor + \sum_{i=1}^p \lfloor a_i/k \rfloor$.
   (d) $a_0 > 1$ *and* $h = 1$ *and* $d = \sum_{i=0}^p a_i - p - 3 - t(k-2)$ *for some* t *where* $p \leq t \leq \lfloor (a_0-1)/k \rfloor + \sum_{i=1}^p \lfloor a_i/k \rfloor$.
   (e) $a_0 > 1$ *and* $h = k-1$ *and* $d = \sum_{i=0}^p a_i - p - 3 - t(k-2)$ *for some* t *where* $p+1 \leq t \leq \sum_{i=0}^p \lfloor a_i/k \rfloor$.

Proof. The first point results from Proposition 6.4 and the fact that

$$[\hat{0},y] \cong Q^{h,k}_{a_0-1}(C_m) \times \Pi^k_{a_1} \times \cdots \times \Pi^k_{a_p}.$$

Definition (6.1) allows us to apply Proposition 6.4 when $a_0 = 1$.

To see that the second point is true, observe that $\tilde{\beta}^d(\hat{0},y) \neq 0$ if and only if there is a nonzero term in the sum in point one. This happens precisely when there is a term in which each factor is nonzero. In other words, we must be able to write $d-2p = r_0+r_1+\cdots+r_p$ where (by Theorem 6.3) we must have $r_i = a_i - 3 - t_i(k-2)$ for some t_i with $1 \leq t_i \leq \lfloor a_i/k \rfloor$ for each $i = 1,\ldots,p$.

If $a_0 = 1$, then $r_0 = -2$ by definition (6.1) and $d = \sum_{i=1}^p a_i - p - 2 - t(k-2)$, where $t = \sum_{i=1}^p t_i$ so that $p \leq t \leq \sum_{i=1}^p \lfloor a_i/k \rfloor$. This is conclusion 2.(a).

Now suppose that $a_0 > 1$ and consider the options for r_0. Suppose $1 < h < k-1$. By Corollary 4.2 we must have either $r_0 = a_0 - 3 - t_0(k-2)$ for some t_0 where $1 \leq t_0 \leq \lfloor (a_0-1)/k \rfloor$ or $r_0 = a_0 - 3 - t_0(k-2) - (h-1)$ for some t_0 where $0 \leq t_0 \leq \lfloor (a_0-h-1)/k \rfloor$. If $h = 1$ or $h = k-1$, then the same corollary gives $r_0 = a_0 - 3 - t_0(k-2)$ with $0 \leq t_0 \leq \lfloor (a_0-1)/k \rfloor$ or with $1 \leq t_0 \leq \lfloor a_0/k \rfloor$, respectively. Substituting for the r_i's, setting $t = \sum_{i=0}^p t_i$, and rearranging, we get the following possibilities:

1. $1 < h < k-1$ and $d = \sum_{i=0}^p a_i - p - 3 - t(k-2)$, where $p+1 \leq t \leq \lfloor (a_0-1)/k \rfloor + \sum_{i=1}^p \lfloor a_i/k \rfloor$.
2. $1 < h < k-1$ and $d = \sum_{i=0}^p a_i - p - 3 - t(k-2) - (h-1)$, where $p \leq t \leq \lfloor (a_0-h-1)/k \rfloor + \sum_{i=1}^p \lfloor a_i/k \rfloor$.
3. $h = 1$ and $d = \sum_{i=0}^p a_i - p - 3 - t(k-2)$, where $p \leq t \leq \lfloor (a_0-1)/k \rfloor + \sum_{i=1}^p \lfloor a_i/k \rfloor$.
4. $h = k-1$ and $d = \sum_{i=0}^p a_i - p - 3 - t(k-2)$, where $p+1 \leq t \leq \sum_{i=0}^p \lfloor a_i/k \rfloor$.

Options 1.–4. correspond to conclusions 2.(b)–(e), respectively. □

Let $\rho^d_{n,m,h,k}$ and $\tilde{\rho}^d_{n,m,h,k}$ denote the ranks of $H^d(M^{h,k}_{n,m})$ and $\tilde{H}^d(M^{h,k}_{n,m})$, respectively.

Theorem 6.6. *Suppose* $1 \leq h < k$ *and* $k > 2$ *and* $n \geq h$.

1. *The groups* $H^d(M^{h,k}_{n,m})$ *are free.*

2.

$$\tilde{\rho}^d_{n,m,h,k} = \sum_{y \in Q^{h,k}_n(C_m)\setminus\hat{0}} \left(\sum_{u_0+\cdots+u_p=d} \tilde{\beta}^{2a_0-4-u_0}_{a_0-1,m,h,k}\ \tilde{\beta}^{2a_1-4-u_1}_{a_1,k} \cdots\ \tilde{\beta}^{2a_p-4-u_p}_{a_p,k} \right).$$

Here a_0 is the size of the zero block of y, $a_1, \ldots, a_p$ are the sizes of the nonzero nonsingleton blocks of y, and the u_i's can assume any integer value.

Proof. The first conclusion follows from the Goresky–MacPherson formula together with the fact that all intervals of an EL-shellable poset are EL-shellable. As for the second conclusion, the Goresky–MacPherson formula and the fact that the (real) codimension[6] of y is $2a_0 + \cdots + 2a_p - 2p - 2$ show that

$$\tilde{\rho}^d_{n,m,h,k} = \sum_{y \in Q^{h,k}_n(C_m)\setminus\hat{0}} \tilde{\beta}^{2a_0+\cdots+2a_p-2p-4-d}(\hat{0}, y).$$

Point 1. of Theorem 6.5 gives

$$\tilde{\rho}^d_{n,m,h,k} = \sum_{y \in Q^{h,k}_n(C_m)\setminus\hat{0}} \left(\sum_{r_0+\cdots+r_p=2a_0+\cdots+2a_p-4p-4-d} \tilde{\beta}^{r_0}_{a_0-1,m,h,k}\ \tilde{\beta}^{r_1}_{a_1,k} \cdots\ \tilde{\beta}^{r_p}_{a_p,k} \right).$$

Making the change of variable $r_i = 2a_i - 4 - u_i$ for $i = 0, 1, \ldots, p$ gives

$$2a_0 + \cdots + 2a_p - 4p - 4 - d = r_0 + r_1 + \cdots + r_p = 2a_0 + \cdots + 2a_p - 4p - 4 - u_0 - \cdots - u_p,$$

whence $d = u_0 + \cdots + u_p$ and

$$\tilde{\rho}^d_{n,m,h,k} = \sum_{y \in Q^{h,k}_n(C_m)\setminus\hat{0}} \left(\sum_{u_0+\cdots+u_p=d} \tilde{\beta}^{2a_0-4-u_0}_{a_0-1,m,h,k}\ \tilde{\beta}^{2a_1-4-u_1}_{a_1,k} \cdots\ \tilde{\beta}^{2a_p-4-u_p}_{a_p,k} \right)$$

as desired. □

Note. We will have $\tilde{\rho}^d_{n,m,h,k} \neq 0$ precisely when one of the summands in the second point is nonzero; each factor in that summand must therefore be nonzero. Theorem 6.3 and Corollary 4.2 show that $\rho^d_{n,m,h,k} \neq 0$ if and only if there exists $y \in Q^{h,k}_n(C_m)$ such that for $i = 1, \ldots, p$ we have $2a_i - 4 - u_i = a_i - 3 - t_i(k-2)$, that is, $u_i = a_i - 1 + t_i(k-2)$, where $1 \leq t_i \leq \lfloor a_i/k \rfloor$, and

- if $h = 1$, then $2a_0 - 4 - u_0 = a_0 - 1 - t_0(k-2) - 2$, that is, $u_0 = a_0 - 1 + t_0(k-2)$, for some t_0 in the range $0 \leq t \leq \lfloor (a_0-1)/k \rfloor$;
- if $1 < h < k-1$, then either
  - $2a_0 - 4 - u_0 = a_0 - 1 - t_0(k-2) - 2$, that is, $u_0 = a_0 - 1 + t_0(k-2)$, for some t_0 in the range $1 \leq t \leq \lfloor (a_0-1)/k \rfloor$ or
  - $2a_0 - 4 - u_0 = a_0 - 1 - t_0(k-2) - (h-1) - 2$, that is, $u_0 = a_0 - 1 + t_0(k-2) + h - 1$, for some t_0 in the range $0 \leq t_0 \leq \lfloor (a_0 - 1 - h)/k \rfloor$;
- if $h = k-1$, then $2a_0 - 4 - u_0 = a_0 - 1 - t_0(k-2) - 2$, that is, $u_0 = a_0 - 1 + t_0(k-2)$, for some t_0 in the range $1 \leq t \leq \lfloor a_0/k \rfloor$.

Taking $t = t_0 + \cdots + t_p$ with $d = u_0 + \cdots + u_p$ shows that $\rho^d_{n,m,h,k} \neq 0$ if and only if there is $y \in Q^{h,k}_n(G)$ so that

[6] The codimension of $y \in Q^{h,k}_n(C_m)$ means the codimension of the corresponding element of $L(\mathcal{B}^{h,k}_{n,m})$.

- if $h = 1$, then $d = \sum_{i=0}^{p} a_i - p - 1 + t(k-2)$ for some t with $p \leq t \leq \lfloor a_0/k \rfloor + \lfloor a_1/k \rfloor + \cdots + \lfloor a_p/k \rfloor$;
- if $1 < h < k-1$, then either
  - $d = \sum_{i=0}^{p} a_i - p - 1 + t(k-2)$ for some t with $p+1 \leq t \leq \lfloor (a_0-1)/k \rfloor + \lfloor a_1/k \rfloor + \cdots + \lfloor a_p/k \rfloor$ or
  - $d = \sum_{i=0}^{p} a_i - p - 1 + t(k-2) + (h-1)$ for some t with $p \leq t \leq \lfloor (a_0-1-h)/k \rfloor + \lfloor a_1/k \rfloor + \cdots + \lfloor a_p/k \rfloor$;
- if $h = k-1$, then $d = \sum_{i=0}^{p} a_i - p - 1 + t(k-2)$ for some t with $p+1 \leq t \leq \lfloor a_0/k \rfloor + \lfloor a_1/k \rfloor + \cdots + \lfloor a_p/k \rfloor$.

Here the a_i's are as in the previous theorem.

REFERENCES

[Ba] H. Barcelo, *On the action of the symmetric group on the free Lie algebra and the partition lattice*, J. Combin. Theory Ser. A, 55 (1990), pp. 93–129.

[Be] N. Bergeron, *A hyperoctahedral analogue of the free Lie algebra*, J. Combin. Theory Ser. A, 58 (1991), pp. 256–278.

[Bj] A. Björner, *Shellable and Cohen-Macaulay partially ordered sets*, Trans. Amer. Math. Soc., 260 (1980), pp. 159–183.

[Bj-Lo-Ya] A. Björner, L. Lovász, and A. Yao, *Linear decision trees: Volume estimates and topological bounds*, in Proceedings of the 24th ACM Symposium on Theory of Computing (May, 1992), ACM Press, New York, 1992, pp. 170–177.

[Bj-Sa] A. Björner and B. Sagan, *Subspace arrangements of type B_n and D_n*, J. Algebraic Combin., 5 (1996), pp. 291–314.

[Bj-Wa1] A. Björner and M. L. Wachs, *Bruhat order of Coxeter groups and shellability*, Adv. Math., 43 (1982), pp. 87–100.

[Bj-Wa2] A. Björner and M. L. Wachs, *On lexicographically shellable posets*, Trans. Amer. Math. Soc., 277 (1983), pp. 323–341.

[Bj-Wa3] A. Björner and M. L. Wachs, *Nonpure shellable complexes and posets* I, Trans. Amer. Math. Soc., 348 (1996), pp. 1299–1327.

[Bj-Wa4] A. Björner and M. L. Wachs, *Nonpure shellable complexes and posets* II, Trans. Amer. Math. Soc., 349 (1997), pp. 3945–3975.

[Bj-We] A. Björner and V. Welker, *The homology of "k-equal" manifolds and related partition lattices*, Adv. Math., 110 (1995), pp. 277–313.

[Bw] A. E. Browdy, *The (Co)homology of Lattices of Partitions with Restricted Block Size*, Ph.D. dissertation, University of Miami, Miami, FL, 1996.

[Ca-Ha-Ro] A. R. Calderbank, P. Hanlon, and R. W. Robinson, *Partitions into even and odd block size and some unusual characters of the symmetric groups*, Proc. London Math. Soc. (3), 53 (1986), pp. 288–320.

[Go-Ma] M. Goresky and R. MacPherson, *Stratified Morse Theory*, Ergeb. Math. Grenzgeb. (3) 14, Springer-Verlag, New York, NY, 1988.

[Go-Wa] E. Gottlieb and M. Wachs, *Cohomology of Dowling lattices and Lie (super)algebras*, Adv. in Appl. Math., 24 (2000), pp. 301–336.

[Ha] P. Hanlon, *The fixed point partition lattices*, Pacific J. Math., 96 (1981), pp. 319–341.

[Ha-Wa] P. Hanlon and M. L. Wachs, *On Lie k-algebras*, Adv. Math., 113 (1995), pp. 206–236.

[Li] S. Linusson, *Partitions with restricted block sizes, Möbius functions, and the k-of-each problem*, SIAM J. Discrete Math., 10 (1997), pp. 18–29.

[Or-So] P. Orlik and L. Solomon, *Combinatorics and topology of hyperplane arrangements*, Invent. Math., 56 (1980), pp. 167–189.

[Qu] D. Quillen, *Homotopy properties of the poset of nontrivial p-subgroups of a group*, Adv. Math., 28 (1978), pp. 101–128.

[St] R. P. Stanley, *Some aspects of groups acting on finite posets*, J. Combin. Theory Ser. A, 32 (1982), pp. 132–161.

[Su1] S. Sundaram, *On the topology of two partition posets with forbidden block sizes*, J. Pure Appl. Algebra, 155 (2001), pp. 271–304.

[Su2] S. Sundaram, *Applications of the Hopf trace formula to computing homology representations*, in Jerusalem Combinatorics '93, Contemp. Math. 178, AMS, Providence, RI, 1994, pp. 277–309.

[Su3] S. Sundaram, *The homology representations of the symmetric group on Cohen-Macaulay subposets of the partition lattice*, Adv. Math., 104 (1994), pp. 225–296.

[Su-Wa] S. Sundaram and M. L. Wachs, *The homology representations of the k-equal partition lattice*, Trans. Amer. Math. Soc., 349 (1997) pp. 935-354.

[Su-We] S. Sundaram and V. Welker, *Group actions on linear subspace arrangements and applications to configuration spaces*, Trans. Amer. Math. Soc., 349 (1997), pp. 1389–1420.

[Wk] J. W. Walker, *Canonical homeomorphisms of posets*, European J. Combin., 9 (1988), pp. 97–107.

[Wa1] M. L. Wachs, *A basis for the homology of the d-divisible partition lattice*, Adv. Math., 117 (1996), pp. 294–318.

[Wa2] M. L. Wachs, *On the (co)homology of the partition lattice and the free Lie algebra*, Discrete Math., 193 (1998), pp. 287–319.

[Wa3] M. L. Wachs, *Whitney homology of semipure shellable posets*, J. Algebraic Combin., 9 (1999), pp. 173–207.

SIAM J. Discrete Math.
Vol. 17, No. 1, pp. 72–87

DETACHMENTS PRESERVING LOCAL EDGE-CONNECTIVITY OF GRAPHS[*]

TIBOR JORDÁN[†] AND ZOLTÁN SZIGETI[‡]

Abstract. Let $G = (V + s, E)$ be a graph with a designated vertex s of degree $d(s)$, and let $f(s) = (d_1, d_2, \ldots, d_p)$ be a partition of $d(s)$ into p positive integers. An $f(s)$-detachment of G is a graph G' obtained by "splitting" s into p vertices, called the pieces of s, such that the degrees of the pieces of s in G' are given by $f(s)$. Thus every edge $sw \in E$ corresponds to an edge of G' connecting some piece of s to w. We give necessary and sufficient conditions for the existence of an $f(s)$-detachment of G in which the local edge-connectivities between pairs of vertices in V satisfy prespecified lower bounds.

Our result is a common generalization of a theorem of Mader on edge splittings preserving local edge-connectivities and a result of Fleiner on $f(s)$-detachments satisfying uniform lower bounds. It implies a conjecture of Fleiner on $f(s)$-detachments preserving local edge-connectivities. By using our characterization we extend a theorem of Frank on local edge-connectivity augmentation of graphs to the case when stars of given degrees are added, and we also solve the local edge-connectivity augmentation problem for 3-uniform hypergraphs.

Key words. detachments of graphs, edge-connectivity, edge splitting

AMS subject classification. 05C40

DOI. 10.1137/S0895480199363933

1. Introduction. We consider loopless undirected graphs which may contain parallel edges. Let $G = (V + s, E)$ be a graph with a designated vertex s. A *degree specification* is a sequence $f(s) = (d_1, d_2, \ldots, d_p)$ of positive integers with $\sum_1^p d_i = d(s)$. An $f(s)$-*detachment* of G is a graph $G' = (V \cup \{s_1, s_2, \ldots, s_p\}, E')$ obtained by "splitting" s into p vertices, $s_1, s_2, \ldots, s_p$, called the *pieces* of s in G', such that the degree of s_i is equal to d_i for $1 \le i \le p$. Thus every edge $sw \in E$ corresponds to an edge of $s_i w \in E'$, connecting some piece of s to w.

Let $H = (W, E)$ be a graph. For two vertices $u, v \in W$ the *local edge-connectivity* between u and v, denoted by $\lambda_H(u, v)$, is the maximum number of pairwise edge-disjoint paths from u to v. Let $U \subseteq W$ and let $r : U^2 \to Z_+$ be a symmetric integer valued function on pairs of vertices of U. We say that r is a *requirement function on* U, and we call H r-*edge-connected in* U if

$$\lambda_H(x, y) \ge r(x, y) \text{ for all pairs } x, y \in U. \tag{1}$$

The main result of this paper is the following necessary and sufficient condition for the existence of an $f(s)$-detachment of G, which is r-edge-connected in V.

Theorem 1.1. *Let $G = (V + s, E)$ be a graph such that there are no cut-edges incident to s. Let $f(s) = (d_1, d_2, \ldots, d_p)$ be a degree specification with $d_i \ge 2$ for all $1 \le i \le p$, and let r be a requirement function on V. Then there exists an $f(s)$-*

[*] Received by the editors November 9, 1999; accepted for publication (in revised form) March 19, 2003; published electronically September 17, 2003. This research was supported in part by Hungarian Scientific Research Fund grants T037547 and F034930 and FKFP grant 0143/2001.

http://www.siam.org/journals/sidma/17-1/36393.html

[†] Department of Operations Research, Eötvös University, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary (jordan@cs.elte.hu).

[‡] Equipe Combinatoire, Université Paris VI, 4, place Jussieu, 75252 Paris, France (szigeti@math.jussieu.fr). This research of this author was supported in part by an Alexander von Humboldt fellowship.

detachment of G, which is r-edge-connected in V, if and only if

$$G \text{ is } r\text{-edge-connected in } V \quad \text{and} \tag{2}$$

$$\lambda_{G-s}(u,v) \geq r(u,v) - \sum_{i=1}^{p} \lfloor d_i/2 \rfloor \tag{3}$$

holds for every pair $u, v \in V$.

The first results on the existence of highly connected detachments (with degree specifications) were obtained by Nash-Williams [8]. The version that we investigate, where the goal is to detach a single vertex s and the degree specification may be arbitrary, was first studied by Fleiner [3]. He proved the following theorem. For a positive integer k we call a graph H *k-edge-connected in U* if H is r-edge-connected in U for the uniform requirement function $r(u,v) \equiv k$, $u, v \in U$.

Theorem 1.2 (see [3]). *Let $G = (V+s, E)$ be k-edge-connected in V for some $k \geq 2$, and let $f(s) = (d_1, d_2, \ldots, d_p)$ be a degree specification with $d_i \geq 2$ for all $1 \leq i \leq p$. Then there exists an $f(s)$-detachment of G which is k-edge-connected in V if and only if $\lambda_{G-s}(u,v) \geq k - \sum_{i=1}^{p} \lfloor d_i/2 \rfloor$ holds for every pair $u, v \in V$.*

Theorem 1.2 follows from Theorem 1.1 by setting $r(u,v) = k$ for every pair $u, v \in V$. Note that Fleiner's result is valid for a family of hypergraphs as well. We discuss this extension and show how it follows from Theorem 1.1 in section 6.

In the same paper Fleiner conjectured that if there are no cut-edges incident to s in $G = (V+s, E)$ and $\lambda_{G-s}(u,v) \geq \lambda_G(u,v) - \sum_{i=1}^{p} \lfloor d_i/2 \rfloor$ holds for every pair $u, v \in V$, then there exists an $f(s)$-detachment G' of G for which $\lambda_{G'}(u,v) = \lambda_G(u,v)$ for every pair $u, v \in V$. Theorem 1.1 implies this conjecture by setting $r(u,v) = \lambda_G(u,v)$ for every $u, v \in V$.

Detachments are closely related to edge splittings. *Splitting off* two edges su, sv in a graph means replacing su, sv by a new edge uv. If $u = v$, then the resulting loop is deleted. Splitting off su, sv in a graph $G = (V+s, E)$ is called *λ-admissible* if $\lambda_{G'}(u,v) = \lambda_G(u,v)$ holds for every pair $u, v \in V$, where G' is the graph obtained by splitting off the pair su, sv. This operation is a useful tool in proving theorems and designing algorithms for connectivity problems. We use and, in Theorem 1.1, we extend the following deep result of Mader on edge splittings preserving local edge-connectivity. The somewhat stronger form of this result that we state here, and a shorter proof, can be found in Frank [4].

Theorem 1.3 (see [7]). *Let $G = (V+s, E)$ be a connected graph with $d(s) \neq 3$ such that there is no cut-edge incident to s. Then there is a λ-admissible splitting at s.*

Proof. We may assume that $d(s) \geq 4$. Let us define a degree specification for s and a requirement function on V by choosing $f(s) = (d_1, d_2)$ with $d_1 = 2$ and $d_2 = d(s) - 2$, and $r(u,v) = \lambda_G(u,v)$ for all pairs $u, v \in V$. We claim that $G, f(s)$ and r satisfy (2) and (3). Clearly, (2) must hold by the definition of r. Furthermore, we have $p = 2$ and $\sum_{i=1}^{p} \lfloor d_i/2 \rfloor = \lfloor d(s)/2 \rfloor$. It is easy to see that $\lambda_{G-s}(u,v) \geq \lambda_G(u,v) - \lfloor d(s)/2 \rfloor$. Thus (3) holds as well. Hence Theorem 1.1 implies that there exists an $f(s)$-detachment $G' = (V \cup \{s_1, s_2\}, E')$ of G which is r-edge-connected in V. This implies that splitting off the edges sx, sy in G, where s_1x, s_1y are the edges incident to s_1 in G', is λ-admissible. □

As a new application, we use Theorem 1.1 to solve the graph augmentation problem where stars of given degrees may be attached to a given graph in order to meet given local edge-connectivity requirements. This extends a theorem of Frank [5] on augmenting the local edge-connectivities by adding a smallest set of new edges. We also solve the local edge-connectivity augmentation problem for 3-uniform hypergraphs.

The organization of the paper is as follows. Section 2 contains further definitions and preliminary results. In section 3 we describe the proof method of our main theorem and show how to reduce the problem to the case when $f(s) = (3, 3, \dots, 3)$. In section 4 we introduce the method of "tight set contraction." By using this method we complete the proof of the main result in section 5. Applications are discussed in section 6.

2. Preliminaries. Let $H = (W, E)$ be a graph. For $X, Y \subseteq W$ we use $d(X, Y)$ to denote the number of edges from $X - Y$ to $Y - X$. Let $d(X) = d(X, V - X)$ denote the *degree* of a set $X \subseteq W$. A singleton set $\{v\}$ is simply denoted by v. Thus $d(v)$ is the degree of vertex $v \in W$. The symbols $\subseteq$ and $\subset$ denote set containment and proper set containment, respectively. Two sets X, Y are said to be *intersecting* if $X \cap Y$, $X - Y$, $Y - X$ are all nonempty.

The degree function of a graph satisfies the following well-known equalities.

PROPOSITION 2.1. *Let $H = (W, E)$ be a graph. For arbitrary subsets $X, Y \subseteq W$,*

$$d(X) + d(Y) = d(X \cap Y) + d(X \cup Y) + 2d(X, Y), \tag{4}$$

$$d(X) + d(Y) = d(X - Y) + d(Y - X) + 2d(X \cap Y, W - (X \cup Y)). \tag{5}$$

Let $H = (W, E)$ be a graph, let $U \subseteq W$ be a subset of vertices with $|U| \geq 2$ and let $r : U^2 \to Z_+$ be a requirement function on U. For set $\emptyset \neq X \subset U$ we define

$$R(X) = \max\{r(u, v) : u \in X, v \in U - X\}. \tag{6}$$

Clearly, R is symmetric on U; that is, $R(X) = R(U - X)$ holds for all $\emptyset \neq X \subset U$. The following property was verified by Frank.

PROPOSITION 2.2 (see [5, Proposition 5.4]). *For any two nonempty subsets X, $Y \subseteq U$, at least one of the following holds:*

$$R(X) + R(Y) \leq R(X \cap Y) + R(X \cup Y), \tag{7}$$

$$R(X) + R(Y) \leq R(X - Y) + R(Y - X). \tag{8}$$

If $X \cup Y = U$, then (8) *always holds (with equality).*

In the rest of this section let $G = (V + s, E)$ be a graph with a designated vertex s and let $r : V^2 \to Z_+$ be a requirement function on V. For sets $\emptyset \neq X \subset V$ we define

$$h(X) = d(X) - R(X). \tag{9}$$

Propositions 2.1 and 2.2 imply the following proposition.

PROPOSITION 2.3. *For any two nonempty subsets $X, Y \subset V$, at least one of the following holds:*

$$h(X) + h(Y) \geq h(X \cap Y) + h(X \cup Y) + 2d(X, Y), \tag{10}$$

$$h(X) + h(Y) \geq h(X - Y) + h(Y - X) + 2d(X \cap Y, V + s - (X \cup Y)). \tag{11}$$

If $X \cup Y = V$, then (11) *always holds (with equality).*

Given a degree specification $f(s) = (d_1, d_2, \dots, d_p)$, we shall use the notation

$$\varphi = \sum_{i=1}^{p} \lfloor d_i/2 \rfloor. \tag{12}$$

It will be convenient to reformulate (2) and (3) in terms of the degree functions of G and $G-s$, functions h and R, and number φ. The following two lemmas are easy to deduce from Menger's theorem.

Lemma 2.4. *Condition* (2) *holds in* $G=(V+s,E)$ *if and only if*

$$h(X)\geq 0 \quad \textit{for every } \emptyset\neq X\subset V. \tag{13}$$

Lemma 2.5. *Condition* (3) *holds in* $G=(V+s,E)$ *if and only if*

$$d_{G-s}(X)\geq R(X)-\varphi \quad \textit{for every } \emptyset\neq X\subset V, \tag{14}$$

which is equivalent to

$$d_G(s,X)\leq h(X)+\varphi \quad \textit{for every } \emptyset\neq X\subset V. \tag{15}$$

We shall need the following operations to construct the required $f(s)$-detachment of G. Let su,sv,sz be distinct edges in $G=(V+s,E)$. The operation 2-*split* (on su,sv) deletes the edges su,sv and adds a new vertex t and two new edges tu,tv. Similarly, operation 3-*split* (on su,sv,sz) deletes su,sv,sz and adds a new vertex t and three new edges tu,tv,tz. Note that (some of) the split edges may be parallel. Let $G=(V+s,E)$ be r-edge-connected in V. We say that a 2-split (or 3-split) is r-*admissible* in G if the resulting graph $G'=(V+s+t,E')$ is also r-edge-connected in V. Let $\emptyset\neq X\subset V$. We say that

$$X \text{ is } \begin{cases} \textit{tight}, \\ \textit{dangerous}, \\ \textit{bad} \end{cases} \quad \text{if} \quad h(X) = \begin{cases} =0, \\ \leq 1, \\ \leq 2. \end{cases}$$

Note that, by definition, tight sets are dangerous, and dangerous sets are bad. For a pair (or triple) of edges su,sv (su,sv,sz) and a set $X\subset V$, we shall use $e(su,sv;X)$ ($e(su,sv,sz;X)$) to denote the number of those edges of the given pair (or triple, respectively) that enter X.

Lemma 2.6. *Let* $G=(V+s,E)$ *be* r*-edge-connected in* V*. Then*

(a) *the* 2*-split on* su,sv *is* r*-admissible if and only if there is no dangerous set* X *with* $e(su,sv;X)=2$*,*

(b) *the* 3*-split on* su,sv,sz *is* r*-admissible if and only if* (i) *there is no tight set* X *with* $e(su,sv,sz;X)\geq 2$ *and* (ii) *there is no bad set* M *with* $e(su,sv,sz;M)=3$.

Proof. We prove only (b). The proof of (a) is similar (but simpler). Let $G'=(V+s+t,E')$ denote the graph obtained from G by a 3-split on su,sv,sz. Observe that for proper subsets $X\subset V$ we have $d_{G'}(X+t)=d_G(X)-(2e(su,sv,sz;X)-3)$.

To see necessity suppose that (i) or (ii) does not hold. Then there is a set $Y\subset V$, which is either tight or bad in G, and for which we have $d_{G'}(Y+t)\leq R(Y)-1$. Thus, for some pair x,y with $x\in Y, y\in V-Y$, and $r(x,y)=R(Y)$, we must have $\lambda_{G'}(x,y)\leq r(x,y)-1$. Hence the 3-split is not r-admissible.

To see sufficiency suppose that the 3-split on su,sv,sz is not r-admissible and let $x,y\in V$ with $\lambda_{G'}(x,y)\leq r(x,y)-1$. Then there is a set $Y\subset V+s+t$ with $x\in Y, y\notin Y$ and $d_{G'}(Y)\leq r(x,y)-1$. Since $d_{G'}$ and r are symmetric, we may assume that $s\notin Y$. Thus, since G is r-edge-connected in V and we have $d_G(N)=d_{G'}(N)$ for all $N\subseteq V$, we must have $t\in Y$. Let $X=Y-t$. Since $d_G(X)\geq R(X)\geq r(x,y)$, we can now deduce that (i) or (ii) does not hold. □

We close this section with two simple lemmas that we shall use in sections 4 and 5. The proofs demonstrate the typical applications of inequalities (10) and (11).

Lemma 2.7. *Let $G = (V + s, E)$ be r-edge-connected in V and let X, Y be intersecting tight sets in G. Then either $X \cap Y$ and $X \cup Y$ are also tight, or $X - Y$ and $Y - X$ are tight and $d(s, X \cap Y) = 0$ holds.*

Proof. We apply Proposition 2.3 to the pair X, Y and use the fact that $h(Z) \geq 0$ for all $\emptyset \neq Z \subset V$ by Lemma 2.4. If (10) holds and $X \cup Y \neq V$, then we have $0 = h(X) + h(Y) \geq h(X \cap Y) + h(X \cup Y) \geq 0$. Thus equality holds everywhere, and hence $X \cap Y$ and $X \cup Y$ are also tight. If (11) holds, then we have $0 = h(X) + h(Y) \geq h(X - Y) + h(Y - X) + 2d(X \cap Y, V + s - (X \cup Y)) \geq 2d(X \cap Y, s)$. This implies $d(s, X \cap Y) = 0$ and $h(X - Y) = h(Y - X) = 0$. Thus $X - Y$ and $Y - X$ are tight. $\square$

We say that a triple su, sv, sz is *semi-admissible* if there is no tight set X with $e(su, sv, sz; X) \geq 2$ (that is, Lemma 2.6(b)(i) holds).

Lemma 2.8. *Let $G = (V + s, E)$ be r-edge-connected in V, let X be a tight set, let su, sv, sz be a semi-admissible triple, and let M be a maximal bad set with $e(su, sv, sz; M) = 3$. Then either $X \subseteq M$, or $d(s, X \cap M) = 0$, $M - X$ is bad, and $e(su, sv, sz; M - X) = 3$.*

Proof. If $X \subseteq M$ or $X \cap M = \emptyset$, then we are done. $M \subseteq X$ cannot hold, since $e(su, sv, sz; M) = 3$ and the triple is semi-admissible. Thus we may assume that X, M is an intersecting pair.

We apply Proposition 2.3 to the pair X, M and use the fact that $h(Z) \geq 0$ for all $\emptyset \neq Z \subset V$ by Lemma 2.4. If (10) holds and $X \cup M \neq V$, then we have $0 + 2 \geq h(X) + h(M) \geq h(X \cap M) + h(X \cup M) \geq 0 + 3$, a contradiction. Here $h(X \cup M) \geq 3$ follows from the maximality of M.

Thus (11) holds. Then we have $0 + 2 \geq h(X) + h(M) \geq h(X - M) + h(M - X) + 2d(X \cap M, V + s - (X \cup M)) \geq 2d(s, X \cap M)$. If $d(s, X \cap M) \geq 1$, then this implies $d(s, X \cap M) = 1$ and $h(M - X) = 0$. Hence $M - X$ is tight, and we must have $e(su, sv, sz; M - X) \geq 2$. This contradicts the fact that su, sv, sz is a semi-admissible triple. Thus $d(s, X \cap M) = 0$ must hold. In this case we get $h(M - X) \leq 2$ and $e(su, sv, sz; M - X) = 3$, as required. $\square$

3. Detachments by admissible and feasible 2-splits. We start this section by an informal description of the main steps of the proof of Theorem 1.1. It will not be difficult to reduce the problem to the case when each term d_i in $f(s)$ is either two or three. Thus either there is a term, say d_p, for which $d_p = 2$ or each term in $f(s)$ is equal to three (in particular, $d_p = 3$). The proof is then by induction on $d(s)$. We shall perform a d_p-split at vertex s (hence creating a new vertex t, and decreasing the degree of s), extend the requirement function to $V + t$, reduce the degree specification by deleting the last term, and then apply induction to show that the resulting graph G' has a good detachment G'' (with respect to the modified requirements and degree specification). From this detachment of G' we can obtain a good $f(s)$-detachment of G by simply adding vertex t to the pieces of s. In order to make this proof work we need to make sure that (a) the d_p-split that we perform preserves r-edge-connectivity in V and (b) the resulting graph satisfies (2) and (3) on vertex set $V + t$ with respect to the extended requirement function and the reduced degree specification. In this section we make this precise and show that the required d_p-split exists when $d_p = 2$. (The other case, when $d_p = 3$, will be settled in section 5.)

In the rest of this section let $G = (V + s, E)$ be a graph satisfying (2) and (3) with respect to a requirement function $r : V^2 \to Z_+$ and degree specification $f(s) = (d_1, d_2, \ldots, d_p)$ with $d_p \in \{2, 3\}$. Furthermore, suppose that r is *smooth*; that is, $r(u, v) \geq 2$ for every pair $u, v \in V$. Note that since r is smooth and (2) holds, it follows that there are no cut-edges in G. Recall that (2) and (3) are equivalent to (13) and (14), respectively, by Lemmas 2.4 and 2.5.

Let $G' = (V + t + s, E')$ be obtained from G by a d_p-split. Here t is the new vertex created by the split. The *extended requirement function* r' on $V + t$ in G' is defined as follows:

$$r'(u, v) = \begin{cases} r(u, v) & \text{if } u, v \in V, \\ 2 & \text{if } u = t, v \in V \text{ (or } v = t, u \in V). \end{cases} \tag{16}$$

The *reduced degree specification* $f'(s)$ is obtained from $f(s)$ by deleting the last term, i.e., $f'(s) = (d_1, \ldots, d_{p-1})$. With the extended requirement function r' we also define function R', defined on the proper subsets X of $V+t$ by $R'(X) = \max\{r'(u, v) : u \in X, v \in (V + t) - X\}$, which satisfies

$$R'(X) = \begin{cases} R(X) & \text{if } \emptyset \neq X \subset V, \\ 2 & \text{if } X = \{t\}, V, \\ R(X - t) & \text{otherwise.} \end{cases} \tag{17}$$

With the reduced degree specification $f'(s)$ we also define the corresponding number $\varphi' = \sum_1^{p-1} \lfloor d_i/2 \rfloor$, which satisfies

$$\varphi' = \varphi - 1. \tag{18}$$

The proof of the next lemma is simple.

LEMMA 3.1. *Suppose that $G' = (V + t + s, E')$ is obtained from G by an r-admissible d_p-split. Then G' is r'-edge-connected in $V + t$.*

Thus by performing an r-admissible d_p-split we can make sure that G' satisfies (2) with respect to r' and $V + t$. To guarantee that G' satisfies (3) as well, we need to find an r-admissible d_p-split satisfying some additional conditions, described in the next lemma.

First observe that condition (3) for G', r', and $f'(s)$ is equivalent to

$$d_{G'-s}(X) \geq R'(X) - \varphi' \text{ for all } \emptyset \neq X \subset V + t. \tag{19}$$

This form of (3) and (18) show that, in order to meet (19), the degree of certain subsets $X \subset V$ in $G' - s$ must be strictly larger than their degree in $G - s$. Let

$$\mathcal{C}' = \{X \subset V : d_{G-s}(X) = R(X) - \varphi\}, \tag{20}$$

and let $\mathcal{C}$ denote the inclusionwise minimal members of $\mathcal{C}'$. Sets in $\mathcal{C}$ are called *cores.*

LEMMA 3.2. *Suppose that $G' = (V + t + s, E')$ is obtained from G by an r-admissible 2-split on su, sv (or 3-split on su, sv, sz). Then G' satisfies* (19) *if and only if $e(su, sv; C) \geq 1$ ($e(su, sv, sz; C) \geq 1$, respectively) for every $C \in \mathcal{C}$.*

Proof. Since the split is r-admissible, r is smooth, and by using (17), we can deduce that (19) holds for $X = \{t\}, V$. For the remaining proper subsets of $V + t$ it follows from the symmetry of R' that (19) is equivalent to

$$d_{G'-s}(X) \geq R(X) - \varphi + 1 \text{ for every } \emptyset \neq X \subset V. \tag{21}$$

The definition of $\mathcal{C}$ shows that (21) holds if and only if there is at least one split edge from s to each of the cores. This proves the lemma. □

We shall also use the fact that $X \in \mathcal{C}'$ holds if and only if

$$d(s, X) = h(X) + \varphi. \tag{22}$$

We say that a 2-split on su, sv is *feasible* if $e(su, sv; C) \geq 1$ for every $C \in \mathcal{C}$. Similarly, a 3-split on su, sv, sz is *feasible* if $e(su, sv, sz; C) \geq 1$ holds for every $C \in \mathcal{C}$. To identify feasible splits the following properties of cores will be useful. We say that a degree specification $f(s) = (d_1, d_2, \dots, d_p)$ is 3*-regular* if $d_i = 3$ for $1 \leq i \leq p$. Note that $d(s) \leq 3\varphi$ and equality holds if and only if $f(s)$ is 3-regular.

Lemma 3.3. *The set of cores satisfies the following:*

(a) *the cores are pairwise disjoint,*

(b) $d(s, C) \geq \varphi$ *for each* $C \in \mathcal{C}$,

(c) $|\mathcal{C}| \in \{0, 2, 3\}$,

(d) *if* $|\mathcal{C}| = 3$, *then* $f(s)$ *is* 3*-regular, each core* C *is tight, and satisfies* $d(s, C) = \varphi$.

Proof. Since G satisfies (3), $G - s$ satisfies (13) with respect to $R - \varphi$, and each core is tight in $G - s$ (also with respect to $R - \varphi$). Thus the minimality of the cores and Lemma 2.7 (applied to $G - s$) imply that (a) holds. Property (b) follows from (22), (2), and Lemma 2.4. To see (c) let C be a core. The symmetry of the degree function of $G - s$ and R in V implies that $V - C \in \mathcal{C}'$. Hence we have $|\mathcal{C}| \geq 2$. It follows from (a), (b), and $d(s) \leq 3\varphi$ that $|\mathcal{C}| \leq 3$. Moreover, $|\mathcal{C}| = 3$ may hold only if $f(s)$ is 3-regular, and $d(s, C) = \varphi$ for each core C. In this case, $h(C) = 0$ follows from (22). Thus C is tight in G. □

Lemma 3.4. *Suppose that* $\mathcal{C} = \{C_1, C_2\}$, $\varphi \geq 2$, *and let* sc_1, sc_2 *be a pair of edges with* $c_i \in C_i$ *for* $1 \leq i \leq 2$. *Then the* 2*-split on* sc_1, sc_2 *is* r*-admissible and feasible.*

Proof. By the choice of c_1, c_2, the 2-split on sc_1, sc_2 is feasible. We shall prove that it is r-admissible as well. Let $\alpha_i = d(s, C_i)$, $1 \leq i \leq 2$. By Lemma 3.3(b) we have $\alpha_i \geq \varphi$, and by (22) we have $h(C_i) = \alpha_i - \varphi$ for $1 \leq i \leq 2$.

For a contradiction suppose that the 2-split is not r-admissible. Then Lemma 2.6 implies that there exists a dangerous set $X \subset V$ with $e(sc_1, sc_2; X) = 2$. Let $\beta_i = d(s, X \cap C_i)$, $1 \leq i \leq 2$. Since $c_i \in X$, we have $\beta_i \geq 1$ for $1 \leq i \leq 2$. If $C_1 \cup C_2 \subseteq X$, then (14) implies $R(X) - \varphi \leq d_{G-s}(X) \leq d(X) - \alpha_1 - \alpha_2 \leq d(X) - 2\varphi \leq R(X) + 1 - 2\varphi$, contradicting the assumption $\varphi \geq 2$. Thus we may assume, without loss of generality, that $C_1 - X \neq \emptyset$.

Claim 3.5. (a) *If* $X \cup C_1 \neq V$, *then* $h(X \cup C_1) \geq \alpha_1 - \varphi + 2$, *and* (b) $h(C_1 - X) \geq \alpha_1 - \varphi + 1 - \beta_1$.

Proof. Since there exist no cores other than C_1 and C_2, each set in $\mathcal{C}'$ contains C_1 or C_2. Thus $V - (X \cup C_1) \notin \mathcal{C}'$ must hold, and by the symmetry of $\mathcal{C}'$, we also have $X \cup C_1 \notin \mathcal{C}'$. Hence (14) implies $R(X \cup C_1) - \varphi + 1 \leq d_{G-s}(X \cup C_1)$. Now, since $\beta_2 \geq 1$, it follows from Lemma 3.3(a) that $d(X \cup C_1) \geq d_{G-s}(X \cup C_1) + \alpha_1 + \beta_2 \geq R(X \cup C_1) + \alpha_1 - \varphi + 2$. Thus (a) holds.

Since C_1 is a core, we have $R(C_1 - X) - \varphi + 1 \leq d_{G-s}(C_1 - X)$. This implies $d(C_1 - X) = d_{G-s}(C_1 - X) + \alpha_1 - \beta_1 \geq R(C_1 - X) - \varphi + 1 + \alpha_1 - \beta_1$. Thus (b) holds. □

Now we apply Proposition 2.3 to X and C_1. If (10) holds and $C_1 \cup X \neq V$, then, by using Lemma 2.4, the fact that X is dangerous, (22), and Claim 3.5(a), we obtain $1 + (\alpha_1 - \varphi) \geq h(X) + h(C_1) \geq h(X \cap C_1) + h(X \cup C_1) \geq \alpha_1 - \varphi + 2$, a contradiction. Thus (11) must hold for X and C_1. Then by Lemma 2.4 and Claim 3.5(b) we have $1 + (\alpha_1 - \varphi) \geq h(X) + h(C_1) \geq h(X - C_1) + h(C_1 - X) + 2d(s, X \cap C_1) \geq \alpha_1 - \varphi + 1 + \beta_1$, a contradiction, since $\beta_1 \geq 1$. This final contradiction shows that the 2-split is r-admissible as claimed. □

Lemma 3.6. *Suppose that* $\varphi \geq 2$ *and* $f(s)$ *is not* 3*-regular. Then there exists an* r*-admissible and feasible* 2*-split in* G.

Proof. By Lemma 3.3(c),(d) we have $|\mathcal{C}| \in \{0, 2\}$. First suppose $\mathcal{C} = \emptyset$. Since

$d(s) \geq 4$ and there are no cut-edges incident to s in G, Theorem 1.3 implies that there is a λ-admissible splitting su, sv at s in G. Since G is r-edge-connected in V, the 2-split on su, sv is r-admissible. Since there are no cores, this 2-split is feasible as well.

Now suppose $|\mathcal{C}| = \{C_1, C_2\}$. In this case, by Lemma 3.3(b), we can choose a pair of edges sc_1, sc_2 with $c_i \in C_i$ for $1 \leq i \leq 2$. By Lemma 3.4 the 2-split on sc_1, sc_2 is r-admissible and feasible. □

4. Contracting tight sets. An important step in the proof of Theorem 1.3 is the contraction of nonsingleton tight sets; see [4, 7]. We shall also follow this approach in section 5. In this section we describe the details of this method and prove that, roughly speaking, contracting nonsingleton tight sets does not change the problem.

Let $G = (V + s, E)$ be a graph and let $r : V^2 \to Z_+$ be a requirement function on V. Suppose that G is r-edge-connected in V and let $T \subset V$ be a tight set in G (with respect to r). Let $\hat{G} = (\hat{V} + s, \hat{E})$ be obtained from G by contracting T into a single vertex t. Each set $\hat{X} \subset \hat{V}$ will correspond to a unique subset of V, which we shall denote by X, defined as follows:

$$X = \begin{cases} \hat{X} & \text{if } t \notin \hat{X}, \\ (\hat{X} - t) \cup T & \text{if } t \in \hat{X}. \end{cases} \tag{23}$$

We shall use $\hat{d}$ and $\hat{\lambda}$ to denote the degree function and the local edge-connectivity function in $\hat{G}$, respectively, and we define a requirement function $\hat{r} : \hat{V}^2 \to Z_+$ for $\hat{G}$ as follows:

$$\hat{r}(u,v) = \begin{cases} r(u,v) & \text{if } u, v \in \hat{V} - t, \\ \max\{r(u,x) : x \in T\} & \text{if } v = t, \\ \max\{r(x,v) : x \in T\} & \text{if } u = t. \end{cases} \tag{24}$$

With $\hat{d}$ and $\hat{r}$ we also define functions $\hat{R}$ and $\hat{h}$ on sets $\hat{X} \subset \hat{V}$ by $\hat{R}(\hat{X}) = \max\{\hat{r}(u,v) : u \in \hat{X}, v \in \hat{V} - \hat{X}\}$ and $\hat{h}(\hat{X}) = \hat{d}(\hat{X}) - \hat{R}(\hat{X})$. The next lemma follows directly from the definition of $\hat{r}$ and the fact that $\hat{d}(\hat{X}) = d_G(X)$ for all $\hat{X} \subset \hat{V}$ (without using the fact that T is tight).

Lemma 4.1. *For $\hat{X} \subset \hat{V}$ we have*

(a) $\hat{R}(\hat{X}) = R(X)$, *and*

(b) $\hat{h}(\hat{X}) = h(X)$.

It follows that $\hat{G}$ is $\hat{r}$-edge-connected in $\hat{V}$. We also obtain the following easy but useful corollaries.

Lemma 4.2. *Suppose that the 2-split on su, sv is r-admissible in G. Then the 2-split on su', sv' is $\hat{r}$-admissible in $\hat{G}$, where su' and sv' denote the edges of $\hat{G}$ corresponding to su and sv.*

Proof. Suppose that su', sv' is not $\hat{r}$-admissible in $\hat{G}$. Then by Lemma 2.6(a) there exists a dangerous set $\hat{X} \subset \hat{V}$ with $\hat{h}(\hat{X}) \leq 1$ and $e(su', sv'; \hat{X}) = 2$. By Lemma 4.1(b) it follows that X is a dangerous set in G with $e(su, sv; X) = 2$. By Lemma 2.6(a) it contradicts the fact that the 2-split on su, sv is r-admissible in G. □

Lemma 4.3. *Suppose that G satisfies* (14) *with respect to r and some degree specification $f(s)$. Then $\hat{G}$ satisfies* (14) *with respect to $\hat{r}$ and $f(s)$.*

Proof. Suppose (14) fails in $\hat{G}$ and let $\hat{W} \subset \hat{V}$ be a set with $d_{\hat{G}-s}(\hat{W}) < \hat{R}(\hat{W}) - \varphi$. Since $d_{G-s}(W) = d_{\hat{G}-s}(\hat{W})$, and by Lemma 4.1(a), we have $d_{G-s}(W) < R(W) - \varphi$, contradicting the fact that (14) holds in G. □

Mader [7] and Frank [4, Claim 3.2] showed that if a splitting is $\hat{\lambda}$-admissible in

$\hat{G}$, then the corresponding splitting is λ-admissible in G. We need a similar fact for 3-splits and arbitrary requirement functions.

LEMMA 4.4. *Suppose that the* 3*-split on* su', sv', sz' *is* $\hat{r}$*-admissible in* $\hat{G}$*. Then the* 3*-split on* su, sv, sz *is* r*-admissible in* G*, where* su, sv, sz *denote the edges corresponding to* su', sv', sz' *in* G*.*

Proof. For a contradiction suppose that the 3-split on su, sv, sz is not r-admissible in G. By Lemma 2.6(b) this implies that there is set $Y \subset V$ in G for which either (i) Y is tight and $e(su, sv, sz; Y) \geq 2$, or (ii) Y is bad and $e(su, sv, sz; Y) = 3$.

CLAIM 4.5. Y *and* T *are intersecting sets.*

Proof. Since T is tight in G, it follows from Lemma 4.1 that t is a singleton tight set in $\hat{G}$ (with respect to $\hat{r}$). Since the 3-split on su', sv', sz' is $\hat{r}$-admissible in $\hat{G}$, it follows from Lemma 2.6(b) that $e(su', sv', sz'; \{t\}) \leq 1$. Thus $Y - T \neq \emptyset$. If $Y \cap T = \emptyset$ or $T \subseteq Y$, then there is a set $\hat{Y}$ in $\hat{G}$ corresponding to Y. By Lemma 4.1(b) we have $h(Y) = \hat{h}(\hat{Y})$, and hence, by Lemma 2.6(b), $\hat{Y}$ shows that the 3-split on su', sv', sz' is not $\hat{r}$-admissible in $\hat{G}$, a contradiction. □

First suppose that Y is tight. In this case Claim 4.5 and Lemma 2.7, applied to Y and T, imply that either $Y \cup T$ is tight, or $Y - T$ is tight and $d(s, Y \cap T) = 0$. In each of these cases we obtain a tight set ($Y \cup T$ or $Y - T$) satisfying (i) which is not intersecting with T. This contradicts Claim 4.5. Thus the 3-split on su, sv, sz is semi-admissible.

Next suppose that Y is bad. We may assume that Y is a maximal bad set. Now Claim 4.5 and Lemma 2.8 imply that $Y - T$ is bad and $e(su, sv, sz; Y - T) = 3$. Thus $Y - T$ satisfies (ii) and is not intersecting with T. This contradicts Claim 4.5 and completes the proof of the lemma. □

5. The proof of Theorem 1.1.

In this section first we consider the case of 3-regular degree specifications and prove that, if the necessary conditions hold, there exists an r-admissible and feasible 3-split in G. This will enable us to complete the proof of our main theorem.

In this section (until the proof of Theorem 1.1) let $G = (V + s, E)$ be a graph satisfying (2) and (3) with respect to a smooth requirement function $r : V^2 \to Z_+$ and 3-regular degree specification $f(s) = (d_1, d_2, \ldots, d_p)$ with $p \geq 2$. These conditions imply that there are no cut-edges in G and $\varphi \geq 2$. Note that (2) and (3) are equivalent to (13) and (14), respectively, by Lemmas 2.4 and 2.5, and recall the definition of cores from section 3 and the fact that G has at most three cores.

First we consider the case when G has three cores. In this case we can find an r-admissible and feasible 3-split without tight set contractions.

LEMMA 5.1. *Suppose that* $\mathcal{C} = \{C_1, C_2, C_3\}$ *and let* sc_1, sc_2, sc_3 *be a triple of edges with* $c_i \in C_i$ *for* $1 \leq i \leq 3$*. Then the* 3*-split on* sc_1, sc_2, sc_3 *is* r*-admissible and feasible.*

Proof. By the choice of c_1, c_2, c_3, the 3-split on sc_1, sc_2, sc_3 is feasible. We shall prove that it is r-admissible as well by showing that conditions (i) and (ii) of Lemma 2.6(b) hold.

For a contradiction first suppose, without loss of generality, that there is a maximal tight set X with $e(sc_1, sc_2; X) = 2$. By Lemma 3.3(d) each core is tight. Thus Lemma 2.7 and the maximality of X imply that $C_1 \cup C_2 \subseteq X$. Now we can use Lemmas 2.5, 3.3(b) and (15) to deduce $\varphi = \varphi + h(X) \geq d(s, X) \geq 2\varphi$. This contradicts the fact that $\varphi \geq 2$. Thus Lemma 2.6(b)(i) holds, and hence the triple is semi-admissible.

Now suppose that there is a maximal bad set M with $e(sc_1, sc_2, sc_3; M) = 3$.

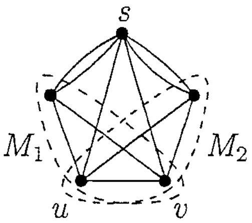


FIG. 1. *An obstacle with respect to the requirement function $r \equiv 4$.*

Since the triple is semi-admissible and each core is tight, Lemma 2.8 implies that $C_1 \cup C_2 \cup C_3 \subseteq M$. By Lemmas 2.5, 3.3(b) and (15), this gives $\varphi + 2 \geq \varphi + h(M) \geq d(s, M) \geq 3\varphi$. Again, this contradicts the fact that $\varphi \geq 2$. Thus Lemma 2.6(b)(ii) holds as well. □

Now we turn to the case when $|\mathcal{C}| \in \{0, 2\}$. In this case we prove the existence of an r-admissible and feasible 3-split by using the method of tight set contractions. To this end, in addition to the assumptions listed at the beginning of this section, in the next three lemmas we shall also assume that

(25) every tight set in G is a singleton.

LEMMA 5.2. *Suppose that* (25) *holds and the* 2*-split on su, sv is r-admissible in G. Then there exists an edge sz for which su, sv, sz is a semi-admissible triple.*

Proof. Let sa be an edge different from su, sv and suppose that the triple su, sv, sa is not semi-admissible. Then there exists a tight set X with $e(su, sv, sa; X) \geq 2$. Since the 2-split on su, sv is r-admissible, Lemma 2.6(a) implies that we must have, without loss of generality, $e(sa, su; X) = 2$. Now (25) implies that $u = a$ and $\{u\}$ is tight. By Lemma 2.5 we have $d(s, u) \leq \varphi$; hence there exists an edge sb, different from sv, with $b \neq u$. If the triple su, sv, sb is not semi-admissible either, then we must have $b = v$ and $\{v\}$ is tight. Moreover, $d(s, v) \leq \varphi$ holds, and hence, since $f(s)$ is 3-regular and hence $d(s) = 3\varphi$, there exists an edge sz with $z \neq u, v$. We conclude that the triple su, sv, sz is semi-admissible, as required. □

We shall strengthen Lemma 5.2 and prove that every r-admissible 2-split is part of an r-admissible 3-split unless G has the following special structure. We say that a graph $H = (W + s, F)$ with $d(s) = 6$ is an *obstacle* (with respect to a given requirement function r on W) if H is r-edge-connected in W and there exist two sets $M_1, M_2 \subset W$ such that

(i) M_1 and M_2 are bad,
(ii) $M_1 - M_2$ and $M_2 - M_1$ are tight, and
(iii) $d_H(s, M_1 - M_2) = d_H(s, M_1 \cap M_2) = d_H(s, M_2 - M_1) = 2$.

An obstacle is shown in Figure 1. It follows from Lemma 2.6(b) that if H is an obstacle, then there is no r-admissible 3-split on su, sv, sz for any choice of the edge sz, where su, sv is the pair of edges from s to $M_1 \cap M_2$. On the other hand, an r-admissible 3-split exists.

LEMMA 5.3. *Suppose that G is an obstacle and* (25) *holds. Then there is an r-admissible* 3*-split in G.*

Proof. Since G is an obstacle and $f(s)$ is 3-regular, we have $d(s) = 6$, $f(s) = (3, 3)$, $\varphi = 2$, and there exist sets $M_1, M_2 \subset V$ satisfying properties (i), (ii), and (iii) above. It follows from (ii) and (25) that $\{a\} = M_1 - M_2$ and $\{b\} = M_2 - M_1$ are singleton tight sets in G. We claim that a 3-split on sa, sb, su is r-admissible, where $u \in M_1 \cap M_2$.

To see this observe that the triple is semi-admissible, since a, b, u are distinct vertices and (25) holds. Furthermore, a bad set M with $e(sa, sb, su; M) = 3$ would satisfy $d(s, M) \geq 5$. This is impossible, since $d(s, M) \leq h(M) + \varphi \leq 4$ by (15). Thus conditions (i) and (ii) of Lemma 2.6(b) hold, and hence the 3-split is r-admissible, as claimed. □

Lemma 5.4. *Suppose that G is not an obstacle,* (25) *holds, and the 2-split on su, sv is r-admissible. Then there exists an edge sz such that the 3-split on su, sv, sz is r-admissible.*

Proof. By Lemma 5.2 the triple su, sv, sa is semi-admissible for some edge sa. If the 3-split on su, sv, sa is r-admissible, then we are done. Otherwise there exists a maximal bad set M with $e(su, sv, sa; M) = 3$ by Lemma 2.6. By (15) we have $d(s, M) \leq h(M) + \varphi \leq 2 + \varphi$. Thus, since $d(s) = 3\varphi$ and $\varphi \geq 2$, it follows that there exists an edge sb with $b \in V - M$. Observe that the triple su, sv, sb is semi-admissible by (25), using the fact that the 2-split on su, sv is r-admissible. Thus either the 3-split on su, sv, sb is r-admissible, or there exists a maximal bad set M' with $e(su, sv, sb; M') = 3$. Clearly, M and M' are intersecting sets.

Apply Proposition 2.3 to the pair M, M'. If (10) holds (and $M \cup M' \neq V$), then we have $2+2 \geq h(M)+h(M') \geq h(M \cap M')+h(M \cup M') \geq 2+3$, a contradiction. Here we used the maximality of M and the fact that the 2-split on su, sv is r-admissible (and hence, by Lemma 2.6(a), $h(M \cap M') \geq 2$ holds). Now suppose that (11) holds, and let $N = V - (M \cup M')$. The existence of the edges su, sv implies that

$$d(M \cap M', N + s) \geq 2. \tag{26}$$

By (11) we have $2 + 2 \geq h(M) + h(M') \geq h(M - M') + h(M' - M) + 2d(M \cap M', V + s - (M \cup M')) \geq 2d(M \cap M', N + s)$. By (26) this implies that $h(M - M') = h(M' - M) = 0$ and $d(M \cap M', N + s) = 2$. Therefore $M - M'$ and $M' - M$ are both tight, and by Lemma 2.5 we have $d(s, M - M'), d(s, M' - M) \leq \varphi$. Suppose that, in addition, we have $d(s, N) = 0$. Then we obtain, by using $\varphi \geq 2$, that $3\varphi \geq 2\varphi + 2 \geq d(s, M - M') + d(s, M' - M) + d(s, M \cap M') + d(s, N) = d(s) = 3\varphi$, which implies that $\varphi = 2$, and G is an obstacle (by choosing $M_1 = M$ and $M_2 = M'$), contradicting our assumption.

Thus $d(s, N) \geq 1$ holds. Let sc be an edge with $c \in N$. As above, it can be seen that su, sv, sc is a semi-admissible triple and either the 3-split on su, sv, sc is r-admissible or there is a maximal bad set M'' with $e(su, sv, sc; M'') = 3$. In the former case we are done. In the latter case we apply Proposition 2.3 to the pairs M, M'' and M', M'', as above, and conclude that $M - M''$, $M'' - M$, $M' - M''$, and $M'' - M'$ are all tight. Since (25) holds and the sets M, M', M'' are pairwise distinct maximal bad sets, it follows that $M \cap M' = M \cap M'' = M' \cap M''$ and $M \cup M' \cup M'' = M \cap M' \cap M'' \cup \{a, b, c\}$.

Claim 5.5. *Let X, Y be disjoint nonempty subsets of V with $d(X, Y) = 0$ and suppose that $X \cup Y$ is bad. Then X and Y are both tight.*

Proof. Clearly, we have $R(X \cup Y) \leq \max\{R(X), R(Y)\}$, so we may suppose, without loss of generality, that $R(X \cup Y) \leq R(X)$ holds. Since G is r-edge-connected in V and r is smooth, Lemma 2.4 implies that $R(X) \leq d(X)$ and $2 \leq R(Y) \leq d(Y)$. Since $d(X, Y) = 0$, we have $d(X) = d(X \cup Y) - d(Y)$. Thus $R(X \cup Y) \leq R(X) \leq d(X) = d(X \cup Y) - d(Y) \leq R(X \cup Y) + 2 - 2$. Hence equality must hold in each of these inequalities, which implies that X and Y are both tight. □

Now we can deduce that $d(c, M'' - c) \geq 1$ as follows. For a contradiction suppose that $d(c, M'' - c) = 0$ holds. Then, by applying Claim 5.5 with $X = \{c\}$ and $Y = M'' - c$, we obtain that the set $M'' - c$ is tight. Since $e(su, sv; M'' - c) = 2$, this contradicts the fact, by Lemma 2.6(a), that the 2-split on su, sv is r-admissible. This proves $d(c, M'' - c) \geq 1$. Thus there is an edge cr with $r \in M'' - c$. We have already shown that $M'' - c = M \cap M'$ and that equality holds in (26). Since the edges su, sv also enter the set $M \cap M'$, we obtain $2 = d(N + s, M \cap M') \geq 3$, a contradiction. This proves the lemma. $\square$

LEMMA 5.6. *There exists an r-admissible and feasible* 3*-split in* G.

Proof. By Lemma 3.3(b),(c) we have $|\mathcal{C}| \in \{0, 2, 3\}$, and there is an edge from s to each of the cores. If $|\mathcal{C}| = 3$, then the lemma follows from Lemma 5.1, so we may assume that $|\mathcal{C}| \in \{0, 2\}$ holds. In this case, starting from G, let us contract nonsingleton tight sets (and update the requirement function) as long as possible, and let $\hat{G} = (\hat{V} + s, \hat{E})$ and $\hat{r}$ denote the resulting graph and requirement function, respectively. By Lemmas 4.1 and 4.3, $\hat{G}$ satisfies (13) and (14) with respect to $\hat{r}$ and $f(s)$, and by construction (25) holds. Note that $\hat{r}$ is smooth, and the degree specification for G and $\hat{G}$ are the same (in particular, $\hat{d}(s) = 3\varphi \geq 6$ holds).

Let us choose a pair su, sv of edges in G in such a way that $e(su, sv; C) \geq 1$ for each core C (if there is any), and the 2-split on su, sv is r-admissible in G. By Theorem 1.3 and Lemma 3.4 this can be done. Lemma 4.2 shows that the 2-split on su', sv' is $\hat{r}$-admissible in $\hat{G}$, where su', sv' are the edges of $\hat{G}$ corresponding to su, sv. Thus, by Lemmas 5.3 and 5.4, either there exists an $\hat{r}$-admissible 3-split su', sv', sz' in $\hat{G}$, or $\hat{G}$ is an obstacle with respect to $\hat{r}$ and there is an $\hat{r}$-admissible 3-split in $\hat{G}$. It follows from Lemma 4.4 that the corresponding 3-splits in G are r-admissible. In the former case the choice of the pair su, sv implies that this 3-split in G is feasible as well. In the latter case, when $\hat{G}$ is an obstacle, (19) can be verified directly by using the fact that the 3-split is r-admissible and $f(s) = (3, 3)$. Thus Lemma 3.2 implies that the 3-split is also feasible. $\square$

We are ready to prove the main result of this paper.

Proof of Theorem 1.1. First we prove necessity. Suppose that $G' = (V + \{s_1, s_2, \ldots, s_p\}, E')$ is an $f(s)$-detachment of G, which is r-edge-connected in V. Since contracting the pieces of s into a single vertex does not decrease the local edge-connectivities between pairs of vertices of V, it follows that G is also r-edge-connected in V. Thus (2) holds. Consider a pair $u, v \in V$ and let $P_1, \ldots, P_l$ be edge-disjoint paths from u to v in G', where $l = \lambda_{G'}(u, v)$. For each piece s_i, $1 \leq i \leq p$, the number of those edges incident to s_i that belong to some of these paths is even. Thus at most $\varphi = \sum_{i=1}^{p} \lfloor d_i/2 \rfloor$ paths can go through the set of pieces of s, and hence at least $l - \varphi$ paths lie entirely in $G' - \{s_1, \ldots, s_p\} = G - s$. Since G' is r-edge-connected in V, this implies $\lambda_{G-s}(u, v) \geq l - \varphi = \lambda_{G'}(u, v) - \varphi \geq r(u, v) - \varphi$, which gives (3).

From now on we prove sufficiency. Suppose that $G = (V + s, E)$, $r : V^2 \to Z_+$, and $f(s) = (d_1, d_2, \ldots, d_p)$ satisfy the hypotheses of the theorem and that (2) and (3) hold. We shall prove that there is an $f(s)$-detachment of G which is r-edge-connected in V. We claim that it is sufficient to consider degree specifications for which $d_i \leq 3$ for $1 \leq i \leq p$. To see this suppose that, without loss of generality, $d_1 \geq 4$, and replace the degree specification $f(s)$ by $f^*(s) = (2, d_1 - 2, d_2, \ldots, d_p)$. The claim follows from the fact that (a) conditions (2) and (3) remain valid (since φ remains the same), and (b) an $f^*(s)$-detachment G^* of G which is r-edge-connected in V gives rise to an $f(s)$-detachment of G, which is also r-edge-connected in V, by identifying two pieces of s in G^* with degree two and degree $d_1 - 2$, respectively. Thus, by iteratively applying this reduction to the degree specification, as long as it is necessary, we may assume that each term in $f(s)$ is equal to two or three. So we can also assume that either

$d_p = 2$, or $d_p = 3$ and $f(s)$ is 3-regular.

Our proof is by induction on $d(s)$. If $d(s) \leq 3$, then, since $d_i \geq 2$ for all $1 \leq i \leq p$, we must have $p = 1$, and the theorem is trivial. Thus we may assume $d(s) \geq 4$. Since $d_i \leq 3$ for all $1 \leq i \leq p$, this implies $p \geq 2$ and $\varphi \geq 2$. By focusing on the 2-edge-connected component of G containing s, we may also assume that G is 2-edge-connected in $V + s$. With this assumption (and since $\varphi \geq 2$) we can increase some of the requirements up to 2, if necessary, without violating conditions (2) and (3). Thus we may also assume $r \geq 2$.

Note that, by Lemmas 2.4 and 2.5, (2) and (3) hold if and only if (13) and (14) hold. If $d_p = 2$, then Lemma 3.6 implies that there is an r-admissible and feasible 2-split in G. If $d_p = 3$ (and hence $f(s)$ is 3-regular), then Lemma 5.6 implies that there is an r-admissible and feasible 3-split in G. Let $G' = (V+t+s, E')$ be the graph obtained from G by an r-admissible and feasible 2- or 3-split, and let r' and $f'(s)$ be the extended requirement function and reduced degree specification, respectively. Now the choice of the split and Lemmas 3.1 and 3.2 imply that G' satisfies (2) and (3) with respect to $V + t$, r', and φ'. Clearly, $d'(s) < d(s)$. Thus, by induction, G' has an $f'(s)$-detachment G'' which is r-edge-connected in V. Since G'' is an $f(s)$-detachment of G, the theorem follows. □

We close this section by extending Theorem 1.1 to arbitrary degree specifications. We need the following definitions. Let $H = (V, E)$ be a graph and let $r : V^2 \to Z_+$ be a requirement function. As before, let $R(X) = \max\{r(u, v) : u \in X, v \in V - X\}$ be defined on sets $X \subset V$, and let $q_H(X) = R(X) - d_H(X)$. A connected component D of H is called a *marginal component* (with respect to r) if $q_H(W) \leq 0$ for every $W \subset D$ and $q_H(D) \leq 1$ hold. Note that if r is smooth, then there are no marginal components in H. We also need a lemma due to Frank.

Lemma 5.7 (see [5, Lemma 5.6]). *Suppose that $H = (V, E)$ has no marginal components, and let $H' = (V + s, E)$ be a graph obtained from H by adding a new vertex s and α new edges between V and s so that H' is r-edge-connected in V. In addition, suppose that γ is an integer satisfying*

$$\sum_{i=1}^{t} q_H(X_i) \leq \gamma \tag{27}$$

for every subpartition $\{X_1, \ldots, X_t\}$ of V. Then it is possible to delete $\alpha - \gamma$ edges incident to s in H' so that none of the remaining new edges is a cut-edge in the resulting graph H'' and that H'' is r-edge-connected in V.

Theorem 5.8. *Let $G = (V + s, E)$ be a graph, let r be a requirement function on V, and suppose that $G - s$ has no marginal components with respect to r. Let $f(s) = (d_1, d_2, \ldots, d_p)$ be an arbitrary degree specification, where the number of $d_i = 1$ terms equals ψ. Then there exists an $f(s)$-detachment of G which is r-edge-connected in V if and only if* (2) *and* (3) *hold, and*

$$\sum_{i=1}^{t} q_{G-s}(X_i) \leq d(s) - \psi \tag{28}$$

holds for every subpartition $\{X_1, \ldots, X_t\}$ of V.

Proof. First we prove the only if direction. Suppose the desired $f(s)$-detachment G' exists. As in the proof of Theorem 1.1, this implies that (2) and (3) hold. Let S be the set of pieces of s in G' and let $L \subseteq S$ denote the set of pieces with degree one. It is easy to see that G' is r-edge-connected in V if and only if $G'-L$ is r-edge-connected in V. Since $G'-L$ is r-edge-connected in V, we have $d_{G'-L}(X) \geq R(X)$ for all $\emptyset \neq X \subset V$. Thus $d_{G'-L}(S-L,X) \geq R(X) - d_{G-s}(X) = q_{G-s}(X)$ for all $\emptyset \neq X \subset V$, which implies (28), since $d_{G'-L}(S-L,V) = d(s) - \psi$.

To see the other direction we can use Lemma 5.7 to show that it is possible to delete ψ edges incident to s in G so that the resulting graph G^* is still r-edge-connected in V and has no cut-edges incident to s. Let $f^*(s)$ be the degree specification obtained from $f(s)$ by deleting each term with $d_i = 1$. It is easy to see that G^* satisfies (3) with respect to r and $f^*(s)$. Now Theorem 1.1 implies that G^* has an $f^*(s)$-detachment $G'' = (V + \{s_1, s_2, \ldots, s_{p-\psi}\}, E'')$ which is r-edge-connected in V. By adding ψ new vertices to G'' and an edge wv_w for some $v_w \in V$ for each new vertex w, we obtain the desired $f(s)$-detachment of G. □

6. Applications and corollaries. In this section we apply Theorem 1.1 to deduce new results on graph and hypergraph augmentation problems. In the local edge-connectivity augmentation problem we are given a graph (or hypergraph) $G = (V,E)$ and a requirement function $r : V^2 \to Z_+$, and the goal is to find a smallest set F of new edges (or hyperedges of given size) for which the augmented graph (or hypergraph) $G' = (V, E+F)$ is r-edge-connected in V. For graphs this problem was solved, in terms of a min-max equality and a polynomial algorithm, by Frank [5]. For hypergraphs it is NP-hard to find a smallest augmentation consisting of hyperedges of size two [2].

First we consider the extension of the graph problem, where, instead of adding edges, we *attach* stars of given degrees. (A star of degree t is the graph $K_{1,t}$.) In other words, we add (an independent set of) new vertices and connect each new vertex s_i to V by d_i edges, where the d_i's are given positive integers. For simplicity, we shall assume that G has no marginal components with respect to r. The following result can be extended to the case when marginal components may exist, following [5]. We leave these details to the interested reader.

THEOREM 6.1. *Let $G = (V,E)$ be a graph and let $r : V^2 \to Z_+$ be a requirement function, such that G has no marginal components with respect to r. Then G can be made r-edge-connected in V by attaching p stars with degrees $d_1, \ldots, d_p$, where $d_i \geq 2$ for all $1 \leq i \leq p$, if and only if*

$$\sum_{i=1}^{t} q_G(X_i) \leq \sum_{i=1}^{p} d_i \tag{29}$$

holds for every subpartition $\{X_1, \ldots, X_t\}$ of V and

$$\lambda_G(u,v) \geq r(u,v) - \sum_{i=1}^{p} \lfloor d_i/2 \rfloor \tag{30}$$

for every pair $u, v \in V$.

Proof. Necessity is easy to see. To prove sufficiency suppose that (29) and (30) hold for G. Observe that the required attachment exists if and only if G can be extended to a graph $G' = (V+s, E')$ by adding a new vertex s such that $d_{G'}(s) = \sum_{i=1}^{p} d_i$, and G' has an $f(s)$-detachment which is r-edge-connected in V for $f(s) = (d_1, d_2, \ldots, d_p)$. Condition (29) and Lemma 5.7 imply that an extension G' which is r-edge-connected in V and which satisfies (2) and $d_{G'}(s) = \sum_{i=1}^{p} d_i$ exists. Condition (30) shows that G' satisfies (3) with respect to r and $f(s)$. Thus Theorem 1.1 implies that G' has an $f(s)$-detachment which is r-edge-connected in V. This proves the theorem. □

With Theorem 6.1 we can also solve the following optimization problem: given G, r, and a positive integer w, determine the smallest number γ_w for which G can be made r-edge-connected by attaching γ_w stars of degree w each. When we augment the local edge-connectivities, attaching stars of degree two is equivalent to adding new edges. Thus for $w = 2$ we obtain Frank's theorem as a corollary. (Note that (29) implies (30) if $d_i = 2$ for $1 \leq i \leq p$.)

Theorem 6.2 (see [5]). *Let $G = (V, E)$ be a graph and let $r : V^2 \to Z_+$ be a requirement function, such that G has no marginal components with respect to r. Then G can be made r-edge-connected by adding γ new edges if and only if $\sum_{i=1}^{t} q(X_i) \leq 2\gamma$ holds for every subpartition $\{X_1, \ldots, X_t\}$ of V.*

A 3-*hypergraph* is a hypergraph with hyperedges of size at most three. The next problem we consider is the local edge-connectivity augmentation problem for 3-hypergraphs, where the new hyperedges that we add are also of size at most three.

Theorem 6.3. *Let $G = (V, E)$ be a 3-hypergraph and let $r : V^2 \to Z_+$ be a smooth requirement function. Then G can be made r-edge-connected by adding γ new hyperedges of size two and β new hyperedges of size three if and only if*

$$\sum_{i=1}^{t} q_G(X_i) \leq 2\gamma + 3\beta \tag{31}$$

holds for every subpartition $\{X_1, \ldots, X_t\}$ of V and

$$\lambda_G(u, v) \geq r(u, v) - (\gamma + \beta) \tag{32}$$

for every pair $u, v \in V$.

Proof. Necessity is easy to see. To see sufficiency, suppose that G satisfies (31) and (32). Let us construct a graph $G' = (V', E')$ by replacing every edge $e = uvw$ of size three in G by a special vertex v_e and three graph edges $v_e u, v_e v, v_e w$. The key observation is that replacing hyperedges of size three by stars of degree three (and vice versa) does not change the local edge-connectivities in V. Thus (32) remains valid in G'. Define a requirement function r' on pairs of vertices of G' by putting $r'(u, v) = r(u, v)$ for all pairs $u, v \in V$ and $r'(x, y) = 0$ otherwise. Since r is smooth, and by the definition of r', it follows that G' has no marginal components with respect to r'. Furthermore, since (31) and (32) hold, and by the definition of G' and r', it can be seen that G' satisfies (29) and (30) with respect to r' and degree sequence $d_1, d_2, \ldots, d_p$, where $p = \gamma + \beta$, $d_i = 2$ for $1 \leq i \leq \gamma$, and $d_i = 3$ for $\gamma + 1 \leq i \leq p$. Thus we can use Theorem 6.1 to deduce that G' can be made r-edge-connected in V by attaching γ stars of degree two and β stars of degree three. Note that, using the fact that $r'(v_e, y) = 0$ if v_e is a special vertex, the proof of Theorem 6.1 shows that the stars can be attached to G' so that no star is attached to any of the special vertices. By replacing each star tu, tv of degree two by an edge uv, and each star tu, tv, tw of degree three by a hyperedge uvw, as well as each star $v_e u, v_e v, v_e w$ by the hyperedge uvw, we obtain an r-edge-connected 3-hypergraph on V, as required. □

Fleiner [3] proved that Theorem 1.2 is valid for 3-hypergraphs as well, provided no hyperedge of size three contains the designated vertex s. In fact, his proof works only in the more general 3-hypergraph setting. He also proved Theorem 6.3 in the special case of uniform requirements. Here we note that the 3-hypergraph version of Theorem 1.2 (and Theorem 1.1) is easy to deduce directly from Theorem 1.1 by the 3-hyperedge/star replacement method that we used in the proof of Theorem 6.3.

The detachment problem considered by Nash-Williams [8, 9] was different in the following sense. Given a graph $G = (V, E)$, a *global degree specification* f for G assigns a degree specification $f(v) = (d_1^v, \dots, d_{p_v}^v)$ to each vertex $v \in V$. We say that G' is a *global f-detachment of* G if G' can be obtained from G by simultaneously detaching each vertex v into p_v pieces such that the degrees of the pieces are given by $f(v)$. Thus every edge uv in G corresponds to an edge connecting some piece of u to some piece of v in G'. Nash-Williams proved the following theorem. (See [1] for a somewhat shorter proof.)

THEOREM 6.4 (see [9]). *Let $G = (V, E)$ be a graph, let f be a global degree specification for G, and let $k \geq 2$ be an integer. Then G has a k-edge-connected global f-detachment if and only if G is k-edge-connected, $d_i^v \geq k$ for every $v \in V$, $1 \leq i \leq p_v$, and neither of the following two cases occurs:*

(a) *k is odd and G has a cut-vertex v with $f(v) = (k, k)$,*

(b) *k is odd, $V = \{u, v\}$, and $f(u) = f(v) = (k, k)$.*

Fleiner [3] showed that Theorem 6.4 can be deduced from Theorem 1.2. It might be interesting to prove some kind of local edge-connectivity version of Theorem 6.4 using Theorem 1.1.

We do not discuss the algorithmic aspects of our results in detail but note that the proofs of Theorem 1.1 and its applications are algorithmic and give rise to polynomial algorithms for constructing an r-edge-connected $f(s)$-detachment, if it exists, and augmenting the local edge-connectivity of a graph or 3-hypergraph optimally by attaching stars or adding hyperedges of size at most three. This follows from the fact that checking whether a 2-split or 3-split is r-admissible and feasible can be done by max-flow computations. An efficient algorithm for the special case of finding λ-admissible splittings is given in [6].

Acknowledgment. This work was done when the first author was a member of BRICS (Basic Research in Computer Science, Centre of the Danish National Research Foundation), Department of Computer Science, University of Aarhus, Denmark, and the second author visited the Research Institute for Discrete Mathematics, Bonn, Germany.

REFERENCES

[1] A. BERG, B. JACKSON, AND T. JORDÁN, *Highly edge-connected detachments of graphs and digraphs*, J. Graph Theory, 43 (2003), pp. 67–77.

[2] B. COSH, B. JACKSON, AND Z. KIRÁLY, *Local Connectivity Augmentation in Hypergraphs is NP-Complete*, manuscript, 2002.

[3] B. FLEINER, *Detachments of vertices of graphs preserving edge-connectivity*, SIAM J. Discrete Math., submitted.

[4] A. FRANK, *On a theorem of Mader*, Discrete Math., 101 (1992), pp. 49–57.

[5] A. FRANK, *Augmenting graphs to meet edge-connectivity requirements*, SIAM J. Discrete Math., 5 (1992), pp. 25–53.

[6] H. N. GABOW, *Efficient splitting off algorithms for graphs*, in Proceedings of the 26th Annual ACM Symposium on the Theory of Computing, ACM, New York, 1994, pp. 696–705.

[7] W. MADER, *A reduction method for edge-connectivity in graphs*, Ann. Discrete Math., 3 (1978), pp. 145–164.

[8] C. ST. J. A. NASH-WILLIAMS, *Detachments of graphs and generalised Euler trails*, in Surveys in Combinatorics 1985, London Math. Soc. Lecture Note Ser. 103, Cambridge University Press, Cambridge, UK, 1985, pp. 137–151.

[9] C. ST. J. A. NASH-WILLIAMS, *Connected detachments of graphs and generalized Euler trails*, J. London Math. Soc. (2), 31 (1985), pp. 17–29.

SIAM J. Discrete Math.
Vol. 17, No. 1, pp. 88–102

CONSTRAINED EDGE-SPLITTING PROBLEMS[*]

TIBOR JORDÁN[†]

Abstract. Splitting off two edges su, sv in a graph G means deleting su, sv and adding a new edge uv. Let $G = (V + s, E)$ be k-edge-connected in V ($k \geq 2$) and let $d(s)$ be even. Lovász proved that the edges incident to s can be split off in pairs in a such a way that the resulting graph on vertex set V is k-edge-connected. In this paper we investigate the existence of such complete splitting sequences when the set of split edges has to meet additional requirements. We prove structural properties of the set of those pairs u, v of neighbors of s for which splitting off su, sv destroys k-edge-connectivity. This leads to a new method for solving problems of this type.

By applying this method we obtain a short proof for a recent result of Nagamochi and Eades on planarity-preserving complete splitting sequences and prove the following new results: let G and H be two graphs on the same set $V + s$ of vertices and suppose that their sets of edges incident to s coincide. Let G (H) be k-edge-connected (l-edge-connected, respectively) in V ($k, l \geq 2$) and let $d(s)$ be even. Then there exists a pair su, sv which can be split off in both graphs preserving k-edge-connectivity in G (l-edge-connectivity in H, respectively), provided $d(s) \geq 6$. If k and l are both even, then such a pair always exists. By using these edge-splitting results and the polymatroid intersection theorem we give a polynomial algorithm for the problem of simultaneously augmenting the edge-connectivity of two graphs by adding a (common) set of new edges of (almost) minimum size.

Key words. edge-splitting, edge-connectivity, graph augmentation, polymatroids, graph algorithms

AMS subject classifications. 05C40, 90C27

DOI. 10.1137/S0895480199364483

1. Introduction.

Edge-splitting is a well-known and useful method for solving problems in graph connectivity. Splitting off two edges su, sv means deleting su, sv and adding a new edge uv. This operation may decrease the edge-connectivity of the graph. The essence of the edge-splitting method is to find a pair of edges which can be split off preserving the edge-connectivity or other connectivity properties of the graph. If such a good pair exists, then one may reduce the problem to a smaller graph which can lead to inductive proofs. Another typical application is the edge-connectivity augmentation problem where splitting off is an important subroutine in some polynomial algorithms. This connection will be discussed in detail in section 5. (For a survey, see [9].)

Let $G = (V + s, E)$ be a graph which is k-edge-connected in V; that is, $d(X) \geq k$ holds for every $\emptyset \neq X \subset V$. Here $d(X)$ denotes the degree of X. Suppose that $d(s)$ is even and $k \geq 2$. Lovász [12] proved that for every edge su there exists an edge sv for which splitting off the pair su, sv preserves k-edge-connectivity in V. We call such a pair *admissible*. By repeated applications of this theorem we can see that all the edges incident to s can be split off in pairs in such a way that the resulting graph

[*]Received by the editors November 19, 1999; accepted for publication (in revised form) March 19, 2003; published electronically September 17, 2003. A preliminary version of this paper appeared in *Proceedings of 7th International IPCO Conference*, Graz, Austria, Lecture Notes in Comput. Sci. 1610, G. Cornuéjols, R. E. Burkard, and G. Woeginger, eds., Springer-Verlag, Berlin, 1999, pp. 273–288. This research was supported in part by the Hungarian Scientific Research Fund grants T037547 and F034930 and FKFP grant 0143/2001.

http://www.siam.org/journals/sidma/17-1/36448.html

[†]Department of Operations Research, Eötvös University, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary (jordan@cs.elte.hu).

(on vertex set V) is k-edge-connected. Such a splitting sequence which isolates s (and preserves k-edge-connectivity in V) is called a *complete (admissible) splitting at s*.

This result gives no information about the structure of the subgraph (V, F) induced by the set F of new edges that we obtain by the splittings (except the degree-sequence of its vertices, which is the same for every complete splitting). Recent problems in edge-connectivity augmentation gave rise to edge-splitting problems, where the goal is to find a complete admissible splitting for which the subgraph of the new edges satisfies some additional requirement. For example, while adding F to $G - s$, one may want to preserve simplicity [2], planarity [13], or bipartiteness [1], too.

The goal of this paper is to develop a new method for solving such "constrained" edge-splitting problems. The basic idea is to define the *nonadmissibility graph* $B(s)$ of $G = (V + s, E)$ on the set of neighbors of s by connecting two vertices x, y if and only if the pair sx, sy is not admissible. We give a complete characterization of those graphs that arise as nonadmissibility graphs. Furthermore, we prove that $B(s)$ is 2-edge-connected if and only if $B(s)$ is a cycle, k is odd, and G has a special structure, which we call round. This structural property turns out to be essential in several edge-splitting problems. Suppose that the additional requirement the split edges have to meet can be given by defining a *constraint graph* $D(s)$ on the neighbors of s in every iteration of the splitting sequence so that the requirement is satisfied if and only if $uv \in E(D(s))$ holds for the admissible pair su, sv that we split off, in every iteration. Clearly, such an admissible pair exists if and only if $D(s)$ is not a subgraph of $B(s)$. Our method is to compare the structure of the graphs $B(s)$ and $D(s)$ that may occur in some iteration. By showing that $D(s)$ can never be the subgraph of the corresponding $B(s)$ we can verify the existence of a complete admissible splitting satisfying the additional requirement.

As a first application of this method we give a simplified proof for a recent result of Nagamochi and Eades [13] on planarity-preserving complete admissible splittings. We also show that some results of Bang-Jensen et al. [1] on partition-constrained complete admissible splittings can be obtained by this method. Then we use our structural results to prove the following "intersection theorem" for admissible splittings: let two graphs $G = (V + s, E)$ and $H = (V + s, K)$ be given for which the sets of edges incident to s in G and H coincide. Let G and H be k- and l-edge-connected in V, respectively $(k, l \geq 2)$. Then there exists a pair of edges su, sv which is admissible in G and H (with respect to k and l, respectively) simultaneously, provided $d(s) \geq 6$. If k and l are both even, then such a pair always exists, and therefore a simultaneously admissible complete splitting exists as well.

By using these edge-splitting results and the polymatroid intersection theorem we give a min-max theorem and a polynomial algorithm for the *simultaneous edge-connectivity augmentation problem*. In this problem two graphs $G' = (V, E)$, $H' = (V, K)$ and two integers $k, l \geq 2$ are given and the goal is to find a smallest set F of new edges for which $G'' = (V, E \cup F)$ and $H'' = (V, K \cup F)$ are k-edge-connected and l-edge-connected, respectively. Our algorithm finds a feasible solution whose size does not exceed the optimum by more than one. If k and l are both even, then the solution is optimal.

1.1. Definitions and notation.

Graphs in this paper are undirected and may contain parallel edges. Let $G = (V, E)$ be a graph. A *subpartition* of V is a collection of pairwise disjoint nonempty subsets of V. A set consisting of a single vertex v is simply denoted by v. An edge joining vertices x and y is denoted by xy. Sometimes xy will refer to an arbitrary copy of the parallel edges between x and y, but this will not cause any confusion. Adding or deleting an edge e in a graph G is denoted by $G+e$ or $G-e$, respectively. The symbols $\subseteq$ and $\subset$ denote set containment and proper set containment, respectively.

For $X, Y \subseteq V$, $d(X,Y)$ denotes the number of edges with one endvertex in $X-Y$ and the other in $Y-X$. We define the *degree* of a subset X as $d(X) = d(X, V-X)$. For example, $d(v)$ denotes the degree of vertex v. The set of *neighbors* of v (or v-*neighbors* for short), that is, the set of vertices adjacent to vertex v, is denoted by $N(v)$. A graph $G=(V,E)$ is *k-edge-connected* if

$$d(X) \geq k \quad \text{for all} \quad \emptyset \neq X \subset V. \tag{1}$$

The operation *splitting off* a pair of edges su, sv at a vertex s means replacing su, sv by a new edge uv. If $u=v$, then the resulting loop is deleted. We use $G_{u,v}$ to denote the graph obtained by splitting off the edges su, sv in G (the vertex s will always be clear from the context). A *complete splitting* at vertex s (with even degree) is a sequence of $d(s)/2$ splittings of pairs of edges incident to s.

2. Preliminaries. The degree function satisfies the following well-known equalities.

Proposition 2.1. *Let $H=(V,E)$ be a graph. For arbitrary subsets $X, Y \subseteq V$,*

$$d(X) + d(Y) = d(X \cap Y) + d(X \cup Y) + 2d(X,Y), \tag{2}$$

$$d(X) + d(Y) = d(X-Y) + d(Y-X) + 2d(X \cap Y, V-(X \cup Y)). \tag{3}$$

In the rest of this section let s be a specified vertex of a graph $G=(V+s,E)$ with degree function d such that $d(s)$ is even and (1) holds with respect to some $k \geq 2$. Saying (1) holds in such a graph G means it holds for all $\emptyset \neq X \subset V$ (in which case G is said to be *k-edge-connected in V*). A set $\emptyset \neq X \subset V$ is called *dangerous* if $d(X) \leq k+1$ and $d(s,X) \geq 2$. (Notice that in the standard definition of dangerous sets property $d(s,X) \geq 2$ is not required.) A set $\emptyset \neq X \subset V$ is *critical* if $d(X)=k$. Two sets $X, Y \subseteq V$ are *crossing* (or X *crosses* Y) if $X-Y$, $Y-X$, $X \cap Y$, and $V-(X \cup Y)$ are all nonempty. Edges sv, st form an *admissible pair* in G if $G_{v,t}$ still satisfies (1). It is well known, and easy to see, that sv, st is not admissible if and only if some dangerous set contains both t and v.

The statements in the following two lemmas can be proved by standard methods using Proposition 2.1. Most of them are well known and appeared explicitly or implicitly in [8], or later in [1] (see also [3]). We omit the proofs.

Lemma 2.2.

(a) *A maximal dangerous set does not cross any critical set.*

(b) *If X is dangerous, then $d(s, V-X) \geq d(s,X)$.*

(c) *If k is even, then two maximal dangerous sets X, Y which are crossing have $d(s, X \cap Y) = 0$.*

Let t be a neighbor of s. A dangerous set X with $t \in X$ is called a *t-dangerous set.*

Lemma 2.3. *Let v be an s-neighbor. Then exactly one of the following holds:*

(i) *The pair sv, su is admissible for every edge $su \neq sv$.*

(ii) *There exists a unique maximal v-dangerous set X.*

(iii) *There exist precisely two maximal v-dangerous sets X, Y. In this case k is odd and we have $d(X) = d(Y) = k+1$, $d(X-Y) = d(Y-X) = d(X \cap Y) = k$, $d(X \cup Y) = k+2$, $d(X \cap Y, V+s-(X \cup Y)) = 1$, $d(s, X-Y) \geq 1$, $d(s, Y-X) \geq 1$, and $d(X \cap Y, X-Y) = d(X \cap Y, Y-X) = (k-1)/2$.*

An s-neighbor v for which (iii) holds in Lemma 2.3 is called *special.*

The previous lemmas include all ingredients of Frank's proof [8] for the next splitting off theorem due to Lovász.

THEOREM 2.4 (see [12, Problem 6.53]). *Suppose that* (1) *holds in* $G = (V + s, E)$*,* $k \geq 2$*,* $d(s)$ *is even, and* $|N(s)| \geq 2$*. Then for every edge* st *there exists an edge* su $(t \neq u)$ *such that the pair* st, su *is admissible.*

3. The structure of nonadmissibility graphs. In this section let $G = (V + s, E)$ be a graph which satisfies (1) with respect to some $k \geq 2$. The *nonadmissibility graph* $B(s) = (N(s), E(B(s)))$ of G (with respect to s) is defined on vertex set $N(s)$. Two vertices $u, v \in N(s)$ $(u \neq v)$ are adjacent in $B(s)$ if and only if the pair su, sv is not admissible in G. Notice that while G may contain parallel edges, $B(s)$ is always a simple graph. It follows from the definition that two edges su, sv $(u \neq v)$ form an admissible pair in G if and only if $uv \in E(\bar{B}(s))$, that is, uv is an edge of the *complement* of $B(s)$.

This notion turns out to be useful in problems where we search for a complete admissible splitting for which the set F of edges obtained by the splittings satisfies some additional property Π. Let $G' := G - s$. For example, $G' + F$ may be required to be simple, planar, or bipartite. If property Π is closed under taking subgraphs, it defines a graph for every iteration of a splitting sequence in the following way. Suppose that by splitting off some admissible pairs we have preserved property Π; that is, $G' + F'$ satisfies Π for the set F' of edges split off so far. Define a *constraint graph* $D_\Pi(s) = (N(s), E(D_\Pi(s)))$ on the set of neighbors of s in such a way that splitting off xy preserves Π (that is, $G' + F' + xy$ satisfies Π) if and only if $xy \in E(D_\Pi(s))$. By definition, a constraint graph is simple. Given Π and its constraint graph $D_\Pi(s)$, an admissible split satisfying Π will be called a $D_\Pi(s)$*-split,* or simply a $D(s)$*-split.* It is clear from the definitions that a $D(s)$-split exists if and only if $D(s)$ and $\bar{B}(s)$ have a common edge. In other words, a $D(s)$-split does not exist if and only if $D(s)$ is a (spanning) subgraph of $B(s)$. It is easy to decide whether a $D_\Pi(s)$-split exists for a given Π. On the other hand, to decide whether a complete admissible splitting satisfying property Π exists may be difficult. (For instance, to decide whether there is a complete admissible splitting in a simple graph that preserves simplicity is NP-complete [11].) Structural properties of the nonadmissibility graph help overcome this difficulty in several cases by the following observation: if $D_\Pi(s)$ cannot be the subgraph of $B(s)$ at any iteration (since, say, it is always connected while the nonadmissibility graph is disconnected), then a complete admissible splitting satisfying property Π exists.

In order to use this kind of argument, we characterize those graphs that arise as nonadmissibility graphs. A vertex v which is adjacent to all the other vertices of the graph is a *dominating vertex.* A complete (subgraph of a) graph will be called a *clique.* The union of two cliques with precisely one vertex in common is a *double clique.* Every double clique has a dominating vertex. In what follows assume that $d(s)$ is even and $|N(s)| \geq 2$. Note that the neighbors of s in some dangerous set of G induce a clique in $B(s)$. The definition of $B(s)$, Lemma 2.3, and Theorem 2.4 imply the following.

LEMMA 3.1. (a) $B(s)$ *has no dominating vertex.* (b) *The neighbors of a vertex* t *in* $B(s)$ *induce a clique unless* t *is special. If* t *is special, the neighbors of* t *in* $B(s)$ *can be covered by two cliques of* $B(s)$*.*

This leads to a simple characterization of nonadmissibility graphs in the case when k is even.

Theorem 3.2. *Suppose that $G = (V+s, E)$ satisfies* (1), *$d(s)$ is even, $|N(s)| \geq 2$, and k is even. Then $B(s)$ is the disjoint union of (at least two) complete graphs.*

Proof. By Lemma 2.3 there is no special s-neighbor in G. Hence, by Lemma 3.1(b), the neighbors of each vertex of $B(s)$ induce a clique. Thus $B(s)$ is the union of pairwise disjoint cliques. $B(s)$ itself cannot be complete by Lemma 3.1(a). □

It is easy to see that every graph consisting of (at least two) disjoint complete graphs can be obtained as a nonadmissibility graph.

Now let us focus on the case when $k \geq 3$ is odd. In this case the complete characterization of nonadmissibility graphs is more complicated. We need the following key lemma.

Lemma 3.3. *Suppose that t is special and let X and Y be the two maximal t-dangerous sets. Let u be a special s-neighbor in $Y - X$. Let Y and Z denote the two maximal u-dangerous sets in G. Then $Z \cap X = \emptyset$, $Y = (X \cap Y) \cup (Z \cap Y)$, and $d(s, Y) = 2$.*

Proof. Notice that for every special vertex $u' \in Y - X$ one of the two maximal u'-dangerous sets must be Y by Lemma 2.3. Thus Y is indeed one of the two maximal u-dangerous sets.

Let v be an s-neighbor in $Z-Y$. First suppose $v \in X-Y$. Since X and Y are the only maximal t-dangerous sets, we have $t \notin Z$. This shows that v is also special and the two maximal v-dangerous sets are X and Z. Lemma 2.3(iii) implies $d(Z) = k+1$ and $d(X-Y) = d(Y-X) = k$. Hence by (3), applied to Z and $X-Y$ and to Z and $Y-X$, we obtain $X - Y \subset Z$ and $Y - X \subset Z$. By Lemma 2.3(iii) we have $d(X \cap Y) = k$. Hence $Z \cap X \cap Y = \emptyset$ by Lemma 2.2(a), provided $Z \cup (X \cap Y) \neq V$. Moreover, $Z \cup (X \cap Y) = V$ implies $d(s, Z) \geq d(s) - 1 > d(s, V - Z)$, using $d(s) \geq 4$. Since Z is dangerous, this would contradict Lemma 2.2(b). Thus we conclude $Z \cap X \cap Y = \emptyset$.

The above verified properties of Z and Lemma 2.3(iii) imply that $k+1 = d(Z) \geq d(X \cap Y, Z) + d(s, Z) \geq d(X \cap Y, X - Y) + d(X \cap Y, Y - X) + 2 = k - 1 + 2 = k+1$. This shows that $d(s, Z) = 2$ and hence $d(s, Z \cup (X \cap Y)) = 3$ holds. We also get $d(Z, V-Z-(X\cap Y)) = 0$, and by Lemma 2.3(iii) we have $d(X\cap Y, V-Z-(X\cap Y)) = 0$ as well. Therefore $d(Z \cup (X \cap Y), V - Z - (X \cap Y)) = 0$ and hence $d(s, Z \cup (X \cap Y)) = d(Z \cup X \cup Y) = 3 \leq k$. This shows $Z \cup X \cup Y$ is dangerous, contradicting the maximality of X.

Thus we may assume that $v \in V - (X \cup Y)$. By Lemma 2.3(iii) we have $d(Z) = k+1$, $d(X-Y) = d(Y-X) = d(X \cap Y) = k$ and there exists an s-neighbor $w \in X - Y$. Clearly, $t \notin Z$ and by the previous argument we may assume $w \notin Z$. We claim that $Z \cap (X - Y) = \emptyset$. Indeed, otherwise Z and $X - Y$ would cross (observe that $t \notin Z \cup (X - Y)$), contradicting Lemma 2.2(a). We claim that $Z \cap (X \cap Y) = \emptyset$ holds as well. This claim follows similarly by Lemma 2.2(a), since $w \notin Z \cup (X \cap Y)$. A third application of Lemma 2.2(a) shows $Y - X \subset Z$. To see this observe that $t \notin Z \cup (Y - X)$ and hence $(Y - X) - Z \neq \emptyset$ would imply that Z and $Y - X$ cross, a contradiction.

Summarizing the previous observations, we obtain $Z \cap X = \emptyset$ and $Y = (X \cap Y) \cup (Z \cap Y)$. By Lemma 2.3(iii) this implies $d(s, Y) = 2$. This proves the lemma. □

A corollary of Lemma 3.3 is the following sharpening of Lemma 3.1(b). Let t be a special s-neighbor and let X and Y be the maximal t-dangerous sets. Then each pair sx, sy with $x \in X - Y$, $y \in Y - X$ is admissible and hence the t-neighbors in $B(s)$ induce two disjoint cliques. (See also [3, section 2].)

We recall some basic definitions and facts. An inclusionwise maximal 2-edge-connected subgraph is a 2-*component.* (By definition, a single vertex is 2-edge-connected.) An edge e is called a *cut-edge* in a connected graph H if $H - e$ is disconnected. It is well known that the 2-components of a graph are pairwise vertex-disjoint and those edges of a connected graph which are not included in any 2-component are precisely the cut-edges. By contracting the 2-components of a connected graph H, we get a tree whose edges correspond to the cut-edges of H. This tree is the 2-*component tree* of H. 2-components corresponding to the leaves of this tree are called *leaves*.

THEOREM 3.4. *Suppose that* $G = (V + s, E)$ *satisfies* (1), $d(s)$ *is even, and* $|N(s)| \geq 2$. *Let* H *be a component of* $B(s)$. *Then* H *is either*

(a) *a clique, or*

(b) *a double clique, or*

(c) *two disjoint cliques connected by a path, or*

(d) *a cycle of length at least four.*

Proof. First we show that each 2-component W of H is either a clique, a double-clique, or a cycle. Suppose W is neither a clique nor a double-clique. Since W is not a clique, there exists a vertex $t \in W$ for which the neighbors of t in W do not form a clique. By Lemma 3.1 and the maximality of W it follows that t is special. Let X and Y denote the two cliques in $B(s)$ corresponding to the two maximal t-dangerous sets of G. By the choice of t and the maximality of W, we have $X \cup Y \subseteq W$. Since $X \cup Y$ forms a double-clique in $B(s)$ and W is connected and is not a double-clique, without loss of generality there is an edge uv in W with $u \in Y - t$, $v \in W - (X \cup Y)$. By applying Lemma 2.3(iii) and Lemma 3.3 to t and u we get $Y = \{t, u\}$ and that u and t have no common neighbors. Let P be a shortest path in $W - tu$ from t to u. Such a path exists since W is 2-edge-connected. By Lemma 3.3 there are no edges in $B(s)$ from $X - t$ to $Y - t$. Thus P has at least three edges and $C := P + tu$ is a cycle of length at least four. We claim that $W = C$. Since P is a shortest path, C has no chords. Therefore each vertex on C is special by Lemma 3.1(b). Suppose there is a vertex $z \in W - C$ that is adjacent to a vertex of C. Since u and t have no common neighbors, we may assume that z and u are nonadjacent. Let $q \neq u$ be the last neighbor of z on P (counting from t to u) and let p be the vertex preceding q on P. (Observe that $q \neq t$ by Lemma 3.1(b). Thus p exists.) By Lemma 3.1(b) and the choice of q, it follows that p and z are adjacent. On the other hand, applying Lemma 3.3 to the special vertex p and its neighbor q, it follows that p and q have no common neighbors, a contradiction. This proves that $W = C$, as required.

It remains to verify that if H is not 2-edge-connected, then it is the union of two disjoint cliques connected by a path. In our argument the following corollaries of Lemma 3.1(b) and Lemma 3.3, respectively, are crucial: (i) $B(s)$ has no "claw" (that is, a vertex that is adjacent to three pairwise nonadjacent vertices); (ii) $B(s)$ contains no induced subgraph isomorphic to the union of a triangle (t, w, u) and two independent edges tx, uy.

Suppose that H is not 2-edge-connected and let Z be a nonsingleton 2-component of H. By our proof above and since $B(s)$ is simple, Z is either a clique (on at least 3 vertices), a double-clique (with each clique containing at least 3 vertices), or a cycle (on at least 4 vertices). Since $Z \neq H$, there is at least one cut-edge e incident to Z. If Z is a double-clique or a cycle, then $Z + e$ has a claw or a bad triangle as in (ii), a contradiction. We have the same conclusion if there are at least two cut-edges incident to Z and Z is a clique. This shows that each nonsingleton 2-component of H is a clique which is also a leaf. Therefore, since $B(s)$ has no claw, the 2-component tree must be a path and hence H is the union of two disjoint cliques, connected by a path, as required. ∎

One can obtain a complete characterization of graphs arising as a nonadmissibility graph (of some graph $G=(V+s,E)$ with $d(s)$ even and $|N(s)|\geq 2$) by extending the properties verified in Lemma 3.1(a) and Theorem 3.4 by some minor observations. For example, if $B(s)$ is disconnected, then it has no C_4 components, and if $B(s)$ consists of two disjoint cliques connected by a path P, then P has an even number of vertices. We leave these details to the interested reader.

Those graphs $G=(V+s,E)$ for which $B(s)$ is 2-edge-connected are of special interest. To describe their structure, first we need some definitions. In a *cyclic partition* $\mathcal{X}=(X_0,\dots,X_{t-1})$ of V the t partition classes $\{X_0,\dots,X_{t-1}\}$ are cyclically ordered. Thus we use the convention $X_t=X_0$, and so on. In a cyclic partition two classes X_i and X_j are *neighboring* if $|j-i|=1$ and *nonneighboring* otherwise. We say that $G'=(V',E')$ is a C_l^p*-graph* for some $p\geq 3$ and some even $l\geq 2$ if there exists a cyclic partition $\mathcal{Y}=(Y_0,\dots,Y_{p-1})$ of V' for which $d'(Y_i)=l$ $(0\leq i\leq p-1)$ and $d'(Y_i,Y_j)=l/2$ for each pair Y_i,Y_j of neighboring classes of $\mathcal{Y}$ (which implies $d'(Y_{i'},Y_{j'})=0$ for each pair of nonneighboring classes $Y_{i'},Y_{j'}$). A cyclic partition of G' with these properties is called *uniform*.

Let $G=(V+s,E)$ satisfy (1) for some odd $k\geq 3$. Such a G is called *round* (from vertex s) if $G-s$ is a $C_{k-1}^{d(s)}$-graph. Note that by (1) this implies that $d(s,V_i)=1$ for each class V_i $(0\leq i\leq d(s)-1)$ of a uniform partition $\mathcal{V}$ of $G-s$.

LEMMA 3.5. *Let $G=(V+s,E)$ satisfy* (1) *for some odd $k\geq 3$. Suppose that G is round from s and let $\mathcal{V}=(V_0,\dots,V_r)$ be a uniform partition of V, where $r=d(s)-1$. Then*

(a) *$G-s$ is $(k-1)$-edge-connected and for every $X\subset V$ with $d_{G-s}(X)=k-1$ either $X\subseteq V_i$ or $V-X\subseteq V_i$ holds for some $0\leq i\leq r$ or $X=\bigcup_i^{i+j}V_i$ for some $0\leq i\leq r, 1\leq j\leq r-1$;*

(b) *for any set I of new edges which induces a connected graph on $N(s)$, the graph $(G-s)+I$ is k-edge-connected;*

(c) *the uniform partition of $G-s$ is unique;*

(d) *$B(s)$ is a cycle on $d(s)$ vertices (which follows the ordering of $\mathcal{V}$).*

Proof. Let $H:=G-s$. Since G satisfies (1) and $d_G(s,V_i)=1$ for $0\leq i\leq r$, we get $d_H(Y)\geq k-1$ if $Y\subseteq V_i$ for some i. Suppose that H is not $(k-1)$-edge-connected and let $X\subset V$ be a maximal set with $d_H(X)\leq k-2$. By the definition of a uniform partition, X cannot be the union of some classes of $\mathcal{V}$. Thus there exists a $V_j\in\mathcal{V}$ for which X and V_j are intersecting. If $X\cup V_j=V$, then $d_H(X)=d_H(V-X)=d_H(V_j-X)\leq k-2$ follows, a contradiction. Thus X and V_j cross. Now (2) and the maximality of X imply $k-1+k-2\geq d_H(V_j)+d_H(X)\geq d(V_j\cap X)+d(V_j\cup X)\geq k-1+k-1$, a contradiction. This proves that H is $(k-1)$-edge-connected. Let $Y\subset V$ satisfy $d_H(Y)=k-1$ and assume that neither Y nor $V-Y$ is a subset of some class of $\mathcal{V}$. If Y crosses some V_j, then by (2) and (3) we have $d_H(Y\cap V_j)=d_H(V_j-Y)=k-1$. This cannot hold by (1) and $d_G(s,V_j)=1$. Thus Y is the union of some classes of $\mathcal{V}$. By the properties of a uniform partition it is clear that these classes have to be consecutive. This proves (a).

Property (a) implies that if I is a set of new edges which induces a connected graph on $N(s)$, then for every $Y\subset V$ with $d_H(Y)=k-1$ there is at least one edge $xy\in I$ with $|Y\cap\{x,y\}|=1$. Thus adding I to H increases the edge-connectivity to at least k. This proves (b).

To see (c) let us fix some s-neighbor v. In a given uniform partition this vertex is the unique s-neighbor in some V_j and by property (a) V_j is the unique maximal set of degree k in G which contains v. This shows that the set of classes of the uniform partitions is unique. By the degree properties of a uniform partition it follows that the cyclic ordering of these classes is also unique. This proves (c).

By definition a uniform partition $\mathcal{V}$ satisfies $d_G(V_i \cup V_{i+1}) = k + 1$ for $0 \leq i \leq r$. Thus $B(s)$ contains a cycle which follows the cyclic ordering of $\mathcal{V}$. This proves (d) when $r = 2$. Suppose $r \geq 3$ and take a dangerous set X in G. By (a) we can see that $d_G(s, X) = 2$ and $d_H(X) = k - 1$ hold and that X must be the union of two consecutive classes of $\mathcal{V}$. Therefore $B(s)$ is a cycle. □

THEOREM 3.6. *Suppose that $G = (V + s, E)$ satisfies* (1) *for some $k \geq 2$, $d(s)$ is even, and $|N(s)| \geq 2$. Then $B(s)$ is 2-edge-connected if and only if $B(s)$ is a cycle of length $d(s)$, k is odd, $d(s) \geq 4$, and G is round from s.*

Proof. Suppose that G is round from s and $d(s) \geq 4$. Lemma 3.5(d) shows $B(s)$ is a cycle and hence $B(s)$ is 2-edge-connected. To see the other direction assume that $B(s)$ is 2-edge-connected. By Theorem 3.4 and Lemma 3.1(a), $B(s)$ is a cycle. Thus, by Lemma 3.1(b), each s-neighbor is special, and hence there are no parallel edges incident to s by Lemma 2.3(iii). This shows that $B(s)$ has $d(s)$ vertices. Let $v_0, \ldots, v_{d(s)-1}$ denote the vertices of $B(s)$, following the cyclic ordering. Let $V_i = X^1_{v_i} \cap X^2_{v_i}$ $(0 \leq i \leq d(s) - 1)$, where $X^1_{v_i}$ and $X^2_{v_i}$ are the two maximal v_i-dangerous sets in G. Now by Lemma 2.3(iii) and Lemma 3.3 it is easy to see that $(V_0, \ldots, V_{d(s)-1})$ is a uniform partition of $G - s$ and that $G - s$ is a $C^{d(s)}_{k-1}$-graph. Hence G is round from s, as required. □

4. Applications.

In this section we apply Theorems 3.2 and 3.6 and give new proofs for previous results from [1] and [13].

4.1. Edge-splitting preserving planarity.

The following theorem is due to Nagamochi and Eades [13]. The new proof we present here seems to be simpler, especially for k even.

THEOREM 4.1 (see [13]). *Let $G = (V + s, E)$ be a planar graph satisfying* (1) *with respect to either an even k or $k = 3$ and suppose that $d(s)$ is even. Then there exists a complete admissible splitting at s for which the resulting graph is planar.*

Proof. Suppose that we are given a fixed planar embedding of G. This embedding uniquely determines a cyclic ordering $\mathcal{C}$ of (the edges incident to s and hence) the neighbors of s. Clearly, splitting off a pair su, sv for which u and v are *consecutive* in this cyclic ordering preserves planarity (and a planar embedding of the resulting graph can be obtained without re-embedding $G - \{su, sv\}$). Thus to see that a complete admissible splitting exists at s which preserves planarity it is enough to prove that $(*)$ there exists an admissible pair su, sv for which u and v are consecutive in $\mathcal{C}$. Let us call such a pair a *consecutive admissible pair.* By repeated applications of $(*)$ we obtain a complete admissible splitting which preserves planarity (and the embedding of $G - s$ as well).

The existence of a consecutive admissible pair can be formulated in terms of a constraint graph. We may assume $|N(s)| \geq 4$. Let $D(s)$ be a cycle defined on the neighbors of s following the cyclic ordering $\mathcal{C}$. Clearly, a consecutive admissible pair exists if and only if there exists a $D(s)$-split. If k is even, then Theorem 3.2 and the fact that $D(s)$ is connected (while the nonadmissibility graph $B(s)$ is disconnected) show that $(*)$ holds in every iteration. (Note that during the process of iteratively splitting off consecutive admissible pairs the set of neighbors as well as the constraint graph $D(s)$ may change. This happens when s loses some neighbor w by splitting off the last copy of the edges sw.) This completes the proof of the theorem in the case when k is even.

Now consider the case $k = 3$. The above argument and Theorem 3.6 show (using the fact that $D(s)$ is 2-edge-connected) that by splitting off consecutive admissible pairs as long as possible, either we find a complete admissible splitting which preserves planarity (and the embedding of $G - s$) or we get stuck in a graph G' which is round from s and for which $B_{G'}(s) = D_{G'}(s)$ holds. In the latter case we need to re-embed some parts of G' in order to complete the splitting sequence and maintain planarity.

Let us consider such a round $G' = (V + s, E')$ with $B_{G'}(s) = D_{G'}(s)$. Let $V_0, \dots, V_{2m-1}$ be the uniform partition of $G' - s$ (where $2m := d_{G'}(s)$) and let v_i be the neighbor of s in V_i $(0 \leq i \leq 2m - 1)$. There exists a face F of $G' - s$ whose boundary includes every s-neighbor. Since $k = 3$ and G' is round, it can be seen that we may assume that F is a finite face in the embedding of $G' - s$ and every edge which connects two consecutive members of the uniform partition of $G' - s$ is on the boundary of F as well as on the boundary of the infinite face. Since G' is round, we can apply Lemma 3.5(a) and deduce that adding the edges v_0v_m and v_iv_{2m-i} $(1 \leq i \leq m - 1)$ to $G' - s$ results in a 3-edge-connected graph G''. Thus G'' can be obtained from G' by a complete admissible splitting. Furthermore, this set of m new edges can be added to the planar embedding of $G' - s$ within face F in such a way that in the resulting embedding of G'' every edge crossing involves the edge v_0v_m. To avoid these edge crossings we can do the following: first we "flip" V_0 and/or V_m in $G' - s$, that is, re-embed the subgraphs induced by V_0 and V_m in such a way that both v_0 and v_m occur on the boundary of the infinite face while the edges leaving V_0 and V_m remain unchanged. Since G' is round, $k = 3$, and v_0 and v_m are s-neighbors in G', it is easy to see that this can be done. Then we can connect v_0 and v_m within the infinite face and add the other new edges as before. This yields a planar embedding of G''. This completes the proof of the theorem. □

The theorem does not hold if $k \geq 5$ is odd; see [13]. Note that the above proof implies that the graphs obtained by a maximal planarity preserving admissible splitting sequence are round for every odd $k \geq 5$. The original proof in [13] also used the flipping operation but in a more sophisticated way. Our proof shows that if $k = 3$, then at most two flippings are sufficient.

4.2. Edge-splitting with partition-constraints. Let $G = (V + s, E)$ be a graph for which (1) holds for some $k \geq 2$ and $d(s)$ is even. Let $\mathcal{P} = \{P_1, P_2, \dots, P_r\}$, $2 \leq r \leq |V|$, be a prescribed partition of V. In order to solve a more general partition-constrained augmentation problem, Bang-Jensen et al. [1] investigated the existence of complete admissible splittings at s for which each split edge connects two distinct classes of $\mathcal{P}$. This problem can also be formulated in terms of constraint graphs, and our results on nonadmissibility graphs can be applied to prove some results from [1]. We briefly sketch this connection below. (Note that the partition-constrained edge-splitting problem turns out to be a special case of the "simultaneous edge-splitting problem" that we discuss in detail in section 5.)

An admissible pair sx, sy is called *allowed* if x and y belong to different classes of $\mathcal{P}$. Let $S := N(s)$, $S_i := S \cap P_i$, and $d_i := d(s, S_i)$. The following definition describes a situation when a complete allowed splitting does not exist if k is odd.

Definition 4.2 (see [1]). *Let $\{A_1, A_2, B_1, B_2\}$ be a partition of V with the following properties in G for some index i, $1 \leq i \leq r$:*

(i) *$d(X) = k$ for $X = A_1, A_2, B_1, B_2$;*

(ii) *$d(X, Y) = 0$ for $(X, Y) = (A_1, A_2), (B_1, B_2)$;*

(iii) *$S \cap X = S_i$ for $X = A_1 \cup A_2$ or $X = B_1 \cup B_2$;*

(iv) *$d_i = d(s)/2$. Such a partition is called a C_4-obstacle in G.*

The following lemma is extracted from [1, Lemmas 2.13, 3.3, 3.4, 5.1].

LEMMA 4.3. *Let $G = (V + s, E)$ be a graph for which* (1) *holds, $d(s)$ is even, and $d_i \leq d(s)/2$ for all $1 \leq i \leq r$. Then*

(i) *if $d(s) \geq 6$, then for any nonempty S_i there is an allowed pair sx, sy with $x \in S_i$;*

(ii) *if $d(s) = 4$, then there exists a complete allowed splitting at s unless k is odd and G contains a C_4-obstacle $\{A_1, A_2, B_1, B_2\}$. In the latter case in graph $G - s$ we have $d(A_1) = d(A_2) = d(B_1) = d(B_2) = k - 1$ and $d(A_1 \cup B_1) = d(A_1 \cup B_2) = k - 1$. Moreover,*

(iii) *if k is even, then there exists a complete allowed splitting at s, and*

(iv) *if k is odd, then there exists a sequence of allowed splittings of length at least $d(s)/2 - 2$.*

Proof. For a nonempty S_i let the constraint graph $D(s)$ be the complete bipartite graph on color classes S_i and $S - S_i$, respectively. Since $d_i \leq d(s)/2$, we have $S - S_i \neq \emptyset$. An allowed pair sx, sy with $x \in S_i$ exists if and only if there exists a $D(s)$-split. $D(s)$ is either 2-edge-connected and is not a cycle, or has a dominating vertex, or is a four-cycle. In the first two cases Theorem 3.2 and Theorem 3.6 show that a $D(s)$-split exists. Moreover, if $D(s)$ is a four-cycle and there is no $D(s)$-split, then G is round and $d(s) = 4$. In that case $G - s$ is a C_4-obstacle with the required properties. This proves (i) and (ii).

Let $d_j \geq d_i$ for every $1 \leq i \leq r$. Splitting off an allowed pair sx, sy with $x \in P_j$ maintains $d_i \leq d(s)/2$ for all $1 \leq i \leq r$. Thus iteratively applying (i) by choosing an S_j with the largest d_j, we can find a complete allowed splitting (if k is even) or a sequence of allowed splittings of length at least $d(s)/2 - 2$. This proves (iii) and (iv). ▯

Note that by Lemma 3.5(c) the C_4-obstacle in (ii), if exists, is unique.

5. Simultaneous edge-splitting and edge-connectivity augmentation. In this section we consider the following optimization problem: let $G = (V, E)$ and $H = (V, K)$ be two graphs on the same set V of vertices and let $k, l \geq 2$ be integers. Find a smallest set F of new edges for which $\hat{G} = (V, E + F)$ is k-edge-connected and $\hat{H} = (V, K + F)$ is l-edge-connected. Let us call this the *simultaneous edge-connectivity augmentation problem.* We give a polynomial algorithm which finds an optimal solution if both k and l are even and finds a solution whose size is at most one more than the optimum, otherwise. One of the two main parts of our algorithm is based on a new splitting off theorem that we prove using Theorems 3.2, 3.4, and 3.6.

If $G = H$ (and $k \geq l$), then our problem reduces to finding a smallest set F of edges for which $\hat{G} = (V, E + F)$ is k-edge-connected. This is the well-solved *k-edge-connectivity augmentation problem.* Several polynomial algorithms are known which can solve this problem optimally. One approach, which is due to Cai and Sun [4] (simplified and extended later by Frank [8]), divides the problem into two parts: first it extends G by adding a new vertex s and a smallest set F' of edges incident to s such that $|F'|$ is even and $G' = (V + s, E + F')$ satisfies (1) with respect to k. Then in the second part, using Theorem 2.4, it finds a complete admissible splitting from s in G'. The resulting set of split edges will be an optimal solution for the k-edge-connectivity augmentation problem; see [8].

We follow and extend this approach for the simultaneous augmentation problem. To do this we have to extend both parts: we need an algorithm which finds a smallest F' incident to s for which $G' = (V + s, E + F')$ and $H' = (V + s, K + F')$ simultaneously satisfy (1) with respect to k and l, respectively, and then we need to verify that there exists a complete splitting at s which is simultaneously admissible in G' and H'.

Both of these extended problems have interesting new properties. While a smallest F' in the first part can be found by a greedy deletion procedure in the k-edge-connectivity augmentation problem, this is not the case in the simultaneous version. Moreover, a complete splitting at s, which is simultaneously admissible in G' and H', does not always exist. (To see this let $V = \{a, b, c, d\}, E = \{ab, bc, cd, da\}, K = \{ac, bd\}, F' = \{sa, sb, sc, sd\}$ and let $k = 3$, $l = 2$.) However, as we shall see, a smallest F' can be found in polynomial time by solving an appropriate "submodular flow" problem. Furthermore, if k and l are both even, then the required complete and simultaneously admissible splitting does exist (and an "almost complete" splitting sequence can always be found).

5.1. Simultaneous edge-splitting. We start with the splitting problem. Let $G = (V + s, E + F)$ and $H = (V + s, K + F)$ be given which satisfy (1) with respect to k and l, respectively. Here F denotes the set of edges incident to s. (For simplicity, we assume $V = V(G) = V(H)$, although for the splitting problem we do not need this.) Suppose that $d(s) := d_G(s) = d_H(s)$ is even. We say that a pair su, sv is *legal* if it is admissible in G as well as in H. A complete splitting sequence at s is *legal* if the resulting graphs (after deleting s) satisfy (1) with respect to k and l, respectively. The property of being legal can be formulated in terms of a constraint graph $D(s)$. Namely, a pair su, sv $(u \neq v)$ is legal if and only if su, sv is a $D(s)$-split in G with respect to $D(s) = \bar{B}_H(s)$. Thus a legal pair exists if and only if $\bar{B}_H(s)$ is not a subgraph of $B_G(s)$.

LEMMA 5.1. *Let $H = (V + s, K + F)$ satisfy* (1) *with respect to some $l \geq 2$ and let $d(s)$ be even. Let D be the complement of the nonadmissibility graph $B_H(s)$ of H. Then one of the following holds:*

(i) *D is 2-edge-connected and D is not a cycle,*

(ii) *D has a dominating vertex,*

(iii) *$D = C_4$,*

(iv) *D arises from a complete bipartite graph $K_{2,m}$ $(m \geq 1)$ by attaching an edge to a vertex of degree m,*

(v) *D consists of two independent edges.*

Proof. Let $B = B_H(s)$ and let $S := N(s)$. If B is 2-edge-connected, then B is a cycle of length $d(s) \geq 4$ by Theorem 3.6. If $d(s) = 4$, then (v) holds, otherwise (i) holds. Thus we may assume that B is not 2-edge-connected.

Case I (B is disconnected). If B has an isolated vertex, then (ii) holds. Otherwise S has a bipartition $S = X \cup Y$, $|X|, |Y| \geq 2$, such that there are no edges from X to Y in B. Let $p := |X|$ and $r := |Y|$. Now D contains a spanning complete bipartite graph $K_{p,r}$. Thus D is 2-edge-connected. If $p = r = 2$, then (iii) holds, otherwise (i) holds.

Case II (B is connected (and has at least one cut-edge)). Now Theorem 3.4 implies that B is the union of two disjoint cliques connected by a path. Lemma 3.1(a) implies that either the connecting path has at least two edges or each of the two cliques has at least two vertices. Thus we can assume that there exists an $X \subset S$ with $|X|, |S - X| \geq 2$ and $d_B(X) = 1$. Let $p := |X|$ and $r := |S - X|$ and let $e = xv$, $v \in X$, be the unique edge leaving X in B. Now D contains a spanning complete bipartite graph $K_{p,r}$ minus one edge. If $p, r \geq 3$, then (i) holds. If $p = r = 2$, then B (as well as D) is a path on four vertices and hence (iv) holds with $m = 1$. Suppose we have $p = 2, r \geq 3$. Let $X = \{v, w\}$. Since B is connected, the edge vw is present in B. Furthermore, $D - x$ is 2-edge-connected. Thus $d_D(x) \geq 2$ implies (i).

If $d_D(x) = 1$, then x is adjacent to every vertex $y \in S - X - \{x\}$ in B. Since v has no neighbors in $S - X - \{x\}$, it follows from Lemma 3.1(b) that x is special and $S - X$ induces a complete graph in B. Thus D arises from a complete bipartite graph $K_{2,m}$ with $m = r - 1 \geq 2$ by attaching an edge (xw) to a vertex of degree m. This gives (iv). □

THEOREM 5.2. *If $d(s) \geq 6$, then there exists a legal pair su, sv. If k and l are both even, then there exists a complete legal splitting at s.*

Proof. Let $D := \bar{B}_H(s)$ and $A := B_G(s)$. If k and l are both even, then Theorem 3.2 shows that D is connected (since it is the complement of a disconnected graph) and A is disconnected. This implies that a legal pair exists for arbitrary even $d(s)$ and hence a complete legal splitting exists as well.

Suppose k is odd and $d(s) \geq 6$. We may assume $|N(s)| \geq 4$. (Otherwise, by Lemma 3.1(a), $B_H(s)$ has an isolated vertex, and hence D has a dominating vertex, while A has no such vertex.) D satisfies one of (i)–(v) in Lemma 5.1. If (i) or (ii) holds, then by Theorem 3.6 and Lemma 3.1(a) D cannot be a (spanning) subgraph of A and hence we are done. If (v) holds, then $B_H(s)$ is cycle of length four. If (iii) holds and D is a subgraph of A, then A is 2-edge-connected and by Theorem 3.6 we have that A is a cycle of length four. In both cases $d(s) = 4$ follows, a contradiction.

Now assume (iv) holds. If $m = 1$, then D (as well as $B_H(s)$) is a path on four vertices. In this case, if D is a subgraph of A, then by Lemma 3.1(a) either A is a four-cycle, in which case $d(s) = 4$ follows, or $A = D$. In the latter case, Lemma 3.1(b) implies that the two inner vertices of A are special in G. Similarly, the two inner vertices of $B_H(s)$ (which are disjoint from the inner vertices of A) are special in H. Therefore by Lemma 2.3(iii) there are no parallel edges incident to s and $d(s) = 4$. This settles case (iv) when $m = 1$.

If (iv) holds with $m \geq 2$ and D is a subgraph of A, then Theorem 3.4 implies that A is a clique with an attached edge. This contradicts Lemma 3.1(b). This completes the proof of the theorem. □

COROLLARY 5.3. *Suppose that $d(s) = 4$ and there exists no legal pair. Then at least one of G and H is round.*

Proof. It follows from the proof of Theorem 5.2 that if no legal pair exists, then $d(s) = 4$ and either one of G and H is round or $B_H(s)$ is a path on four vertices. We show the latter case is impossible. Let the path be (a, b, c, d). By the definition of $B_H(s)$, there exists a dangerous set X in H which contains b and c. Since $d(s) = |N(s)| = 4$, Lemma 2.2(b) implies $d(s, X) = d(s, V - X) = 2$. Therefore $d_H(X) = d_H(V - X)$, $V - X$ is also dangerous, and $a, d \in V - X$. In this case a and d should also be adjacent in $B_H(s)$, a contradiction. □

Note that a complete splitting sequence which is simultaneously admissible in three (or more) graphs does not necessarily exist, even if each of the edge-connectivity values is even.

We remark that the partition-constrained splitting problem can be reduced to a simultaneous edge-splitting problem where at least one of k and l is even. To see this suppose that an instance of the partition-constrained splitting problem is given as in the beginning of section 4.2, satisfying $d_m \leq d(s)/2$, where $d_m := \max_i\{d_G(s, P_i)\}$ in $G = (V + s, E + F)$. Let $S := N_G(s)$. Build graph $H = (S + x + s, K + F)$ as follows. For each set $S \cap P_i$ in G let the corresponding set in H induce a $(2d_m)$-edge-connected graph (say, a complete graph with sufficiently many parallel edges or a singleton). The edges incident to s in G and H coincide. Then from vertex x of H add $2d_m - d_G(s, P_i)$ parallel edges to some vertex of $S \cap P_i$ $(1 \leq i \leq r)$. Now H

satisfies (1) with respect to $l := 2d_m$. It is easy to see that a complete admissible splitting satisfying the partition-constraints in G exists if and only if there exists a complete legal splitting in the pair G, H. This shows that characterizing the pairs G, H for which a complete legal splitting does not exist (even if one of k and l is even) is at least as difficult as the solution of the partition-constrained problem [1].

5.2. Simultaneous edge-connectivity augmentation. Let $G = (V, E)$ and $H = (V, K)$ be two graphs on the same set of vertices. First we show how to find a smallest F' for which the extended graphs $G' = (V+s, E+F')$ and $H' = (V+s, K+F')$ simultaneously satisfy (1) with respect to k and l, respectively. We need some results from the theory of polymatroids.

Let V be a finite ground-set and let $p : 2^V \to \mathbb{Z} \cup \{-\infty\}$ be an integer valued function for which $p(\emptyset) = 0$. We call p *fully supermodular* if $p(X) + p(Y) \leq p(X \cap Y) + p(X \cup Y)$ holds for every $X, Y \subseteq V$. A function $p : 2^V \to \mathbb{Z} \cup \{-\infty\}$ is *skew supermodular* if for every $X, Y \subseteq V$ either the above submodular inequality holds or $p(X) + p(Y) \leq p(X - Y) + p(Y - X)$. If $p(Y) \leq p(X)$ holds for every $Y \subseteq X$, then p is *monotone.* For a fully supermodular and monotone function p the set $C(p) := \{x \in \mathbb{R}^V : x \geq 0, x(A) \geq p(A) \text{ for every } A \subseteq V\}$ is called the *contra-polymatroid* of p. The next result is due to Frank.

THEOREM 5.4 (see [8]). *Let p be a skew supermodular function. Then $C(p)$ is a contra-polymatroid whose unique (monotone, fully supermodular) defining function $\bar{p}$ is given by*

$$\bar{p}(X) := \max\left(\sum_{i=1}^{t} p(X_i) : \{X_1, \ldots, X_t\} \textit{ is a subpartition of } X\right).$$

Given a graph $G = (V, E)$ and $k \in \mathbb{Z}_+$, let $p_G^k : 2^V \to \mathbb{Z}$ be defined by $p_G^k(X) := k - d_G(X)$ if $\emptyset \neq X \neq V$ and $p_G^k(\emptyset) = p_G^k(V) = 0$. This function p_G^k is skew supermodular by Proposition 2.1. Following [8], we say that a vector $z : V \to \mathbb{Z}_+$ is an *augmentation vector* of G (with respect to k) if $z(X) \geq p_G^k(X)$ for every $\emptyset \neq X \subseteq V$. Observe that $G' = (V + s, E + F')$ satisfies (1) with respect to k if and only if $z(v) := d_{F'}(v)$ $(v \in V)$ is an augmentation vector. Hence by Theorem 5.4 the problem of finding a smallest F' for which $G' = (V + s, E + F')$ satisfies (1) can be reduced to finding an integer valued element of the contra-polymatroid $C(p_G^k)$ for which $z(V)$ is minimum. This can be done by a greedy algorithm [8]. Similarly, adding F' is simultaneously good for G and H if and only if $z(X) \geq \max\{p_G^k(X), p_H^l(X)\}$ for every $\emptyset \neq X \subseteq V$, where $z(v) := d_{F'}(v)$ $(v \in V)$. Let us call such a z a *common augmentation vector* of G and H. Clearly, finding the smallest F' for which $G' = (V + s, E + F')$ and $H' = (V + s, K + F')$ satisfy (1) with respect to k and l, respectively, can be solved by finding an integer valued $z \in C(p_G^k) \cap C(p_H^l)$ for which $z(V)$ is minimum. This problem is also tractable by Edmonds' polymatroid intersection theorem, which provides the following min-max equality (see also [7, 10] and [15, Corollary 46.1c]).

THEOREM 5.5 (see [6]). *Let p_1 and p_2 be monotone, fully supermodular functions on V with $p_1(\emptyset) = p_2(\emptyset) = 0$. Then*

$$\min\{z(V) : z \in \mathbb{Z}^V, z \in C(p_1) \cap C(p_2)\} = \max\{p_1(T) + p_2(V - T) : T \subseteq V\}. \tag{4}$$

There exist efficient algorithms for finding a minimizer in (4). In fact, this problem is a special "submodular flow" problem, and the weighted case is also solvable by an algorithm of Cunningham and Frank [5]. See Schrijver [15, Chapter 47] for more details on the algorithmic aspects. Thus one can find a smallest common augmentation vector in polynomial time.

Summarizing our observations we obtain the following algorithm for the simultaneous edge-connectivity augmentation problem. Let $G = (V, E)$ and $H = (V, K)$ be the pair of input graphs.

Step 1. Find a common augmentation vector for G and H for which $z(V)$ is as small as possible.

Step 2. Add a new vertex s to each of G and H and $z(v)$ parallel edges from s to v for every $v \in V$. If $z(V)$ is odd, then add one more edge sw for some $w \in V$.

Step 3. Find a maximal legal splitting sequence $\mathcal{S}$ at s in the resulting pair of graphs. If $\mathcal{S}$ is complete, let the solution F consist of the set of split edges. Otherwise splitting off $\mathcal{S}$ results in a pair of graphs G', H' for which $d(s) = |N(s)| = 4$. In this case let the solution F be the union of the split edges and a set I of three properly chosen additional edges that form a path on the four s-neighbors.

The following theorem shows the correctness of the above algorithm and proves that the solution set F is (almost) optimal. Let us define

$$\Phi_{k,l}(G,H) = \max\left\{\sum_{1}^{r}(k - d_G(X_i)) + \sum_{r+1}^{t}(l - d_H(X_i) :\right.$$
$$\left.\{X_1, \dots, X_t\} \text{ is a subpartition of } V; 0 \le r \le t\right\}.$$

The size of a smallest simultaneous augmenting set for G and H (with respect to k and l, respectively) is denoted by $OPT_{k,l}(G, H)$.

THEOREM 5.6. $\lceil\Phi_{k,l}(G,H)/2\rceil \le OPT_{k,l}(G,H) \le \lceil\Phi_{k,l}(G,H)/2\rceil + 1$. *If k and l are both even, then $OPT_{k,l}(G,H) = \lceil\Phi_{k,l}(G,H)/2\rceil$ holds.*

Proof. It is easy to see that $\lceil\Phi_{k,l}(G,H)/2\rceil \le OPT_{k,l}(G,H)$ holds. We shall prove that the above algorithm results in a simultaneous augmenting set F with size at most $\lceil\Phi_{k,l}(G,H)/2\rceil + 1$ (and with size $\lceil\Phi_{k,l}(G,H)/2\rceil$ if k and l are both even). It follows from Theorems 5.4 and 5.5, and our remarks on common augmentation vectors, that for the vector z that we obtain in Step 1 of the above algorithm we have $z(V) = \Phi_{k,l}(G,H)$. Hence we have $2\lceil\Phi_{k,l}(G,H)/2\rceil$ edges incident to s at the end of Step 2. By Theorem 5.2 we can find a maximal sequence of legal splittings in Step 3 which is either complete or results in a pair of graphs G', H', where $d(s) = |N(s)| = 4$. In the former case the set F of split edges, which is clearly a feasible simultaneous augmenting set, has size $\lceil\Phi_{k,l}(G,H)/2\rceil$ and hence is optimal. If k and l are both even, then such a complete legal splitting always exists, proving $OPT_{k,l}(G,H) = \lceil\Phi_{k,l}(G,H)/2\rceil$. In the latter case one of G' and H', say G', is round by Corollary 5.3. There exists a complete admissible splitting in H' by Theorem 2.4. Let $e = uv, f = xy$ be the two edges obtained by such a complete splitting. Let $g = vx$. Adding e and f to $H' - s$ gives an l-edge-connected graph, and by Lemma 3.5(b) adding the edge set $I := \{e, f, g\}$ to $G' - s$ yields a k-edge-connected graph. Thus the set of edges F, which is the union of the edges obtained by the maximal legal splitting sequence and the edge set I, is a simultaneous augmenting set. We also have $|F| = \lceil\Phi_{k,l}(G,H)/2\rceil + 1$, as required. □

There are examples showing $OPT = \lceil\Phi/2\rceil + 1$ may hold. (Take $V = \{a, b, c, d\}$, $E = \{ab, bc, cd, da\}$, $K = \{ac, bd\}$ and let $k = 3$, $l = 2$.) It is easy to see that the above algorithm can be implemented in polynomial time. As we pointed out, Step 1 can be solved in polynomial time. One approach to solve Step 3 efficiently is using maximum flow computations to check whether a pair of edges is legal or not. We omit the details.

Since the weighted version of Step 1 is also tractable, extensions of the simultaneous augmentation problem to "vertex-induced" cost functions can also be solved (see [8] for such generalizations of the k-edge-connectivity augmentation problem). We also note that in a recent paper Nagamochi and Ibaraki [14, Corollary 1] gave an efficient algorithm for Step 1 which is based on minimum cost flow computations.

Acknowledgments. I thank Bill Jackson, Tamás Király, and the anonymous referees for several useful comments. This work was done when the author was a member of BRICS (Basic Research in Computer Science, Centre of the Danish National Research Foundation), Department of Computer Science, University of Aarhus, Denmark.

REFERENCES

[1] J. Bang-Jensen, H. N. Gabow, T. Jordán, and Z. Szigeti, *Edge-connectivity augmentation with partition constraints*, SIAM J. Discrete Math., 12 (1999), pp. 160–207.

[2] J. Bang-Jensen and T. Jordán, *Edge-connectivity augmentation preserving simplicity*, SIAM J. Discrete Math., 11 (1998), pp. 603–623.

[3] J. Bang-Jensen and T. Jordán, *Splitting off edges within a specified subset preserving the edge-connectivity of the graph*, J. Algorithms, 37 (2000), pp. 326–343.

[4] G. R. Cai and Y. G. Sun, *The minimum augmentation of any graph to a k-edge-connected graph*, Networks, 19 (1989), pp. 151–172.

[5] W. H. Cunningham and A. Frank, *A primal-dual algorithm for submodular flows*, Math. Oper. Res., 10 (1985) pp. 251–262.

[6] J. Edmonds, *Submodular functions, matroids, and certain polyhedra*, in Combinatorial Structures and Their Applications, Gordon and Breach, New York, 1970, pp. 69–87.

[7] A. Frank, *Generalized polymatroids*, in Finite and Infinite Sets, A. Hajnal et al., eds., North–Holland, Amsterdam, 1984, pp. 285–294.

[8] A. Frank, *Augmenting graphs to meet edge-connectivity requirements*, SIAM J. Discrete Math., 5 (1992), pp. 22–53.

[9] A. Frank, *Connectivity augmentation problems in network design*, in Mathematical Programming: State of the Art 1994, J. R. Birge and K. G. Murty, eds., The University of Michigan, Ann Arbor, MI, 1994, pp. 34–63.

[10] A. Frank and É. Tardos, *Generalized polymatroids and submodular flows*, Math. Program., 42 (1988), pp. 489–563.

[11] T. Jordán, *Two NP-Complete Augmentation Problems*, preprint 8/1997, IMADA, Syddansk Universitet, Odense, Denmark; available online from http://www.imada.sdu.dk/Research/Preprints/.

[12] L. Lovász, *Combinatorial Problems and Exercises*, North–Holland, Amsterdam, 1979.

[13] H. Nagamochi and P. Eades, *Edge-splitting and edge-connectivity augmentation in planar graphs*, in Proceedings of the 6th International IPCO Conference, Houston, TX, Lecture Notes in Comput. Sci. 1412, R. E. Bixby, A. E. Boyd, and R. Z. Rios-Mercado, eds., Springer-Verlag, Berlin, 1998, pp. 96–111; *An edge-splitting algorithm in planar graphs*, J. Comb. Optim., 7 (2003), pp. 137–159.

[14] H. Nagamochi and T. Ibaraki, *Polyhedral structure of submodular and posi-modular systems*, Discrete Appl. Math., 107 (2000), pp. 165–189.

[15] A. Schrijver, *Combinatorial Optimization—Polyhedra and Efficiency*, Springer-Verlag, Berlin, 2003.

SIAM J. Discrete Math.
Vol. 17, No. 1, pp. 103–113

ON Z_{2^k}-LINEAR AND QUATERNARY CODES*

H. TAPIA-RECILLAS[†] AND G. VEGA[‡]

Abstract. For any integer $k \geq 1$, an isometry between codes over $\mathbb{Z}_{2^{k+1}}$ and codes over $\mathbb{Z}_4$ is defined and used to give an equivalent generalization of the Gray map to the one introduced in [C. Carlet, *IEEE Trans. Inform. Theory*, 44 (1998), pp. 1543–1547]. Several results related to the linearity or nonlinearity of codes over $\mathbb{Z}_{2^{k+1}}$, as well as its corresponding images under this map, are given. These results are similar to those presented in Theorems 4, 5, and 6 of [A. R. Hammons, Jr., P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Solé, *IEEE Trans. Inform. Theory*, 40 (1994), pp. 301–319] for codes over $\mathbb{Z}_4$.

Key words. linear codes, Gray map, finite rings

AMS subject classifications. 94B05, 11T71

DOI. 10.1137/S0895480101397219

1. Introduction.

For a given integer $n \geq 1$ the Gray map is a bijective function from $\mathbb{Z}_4^n$ into $\mathbb{F}_2^{2n}$, and its main quality is that it is an isometry; in other words, it is a distance preserving function with respect to the Lee and Hamming distances. With this property in mind, the Gray map has been used extensively to construct binary codes from quaternary codes (i.e., codes over $\mathbb{Z}_4$). In [4] the authors introduce a new kind of binary code called a $\mathbb{Z}_4$-linear code, establishing that a binary code of even length is $\mathbb{Z}_4$-linear if its coordinates can be arranged so that it is the image under the Gray map of a quaternary linear code. Furthermore, the authors give necessary and sufficient conditions for a binary code to be $\mathbb{Z}_4$-linear, as well as conditions under which the binary Gray map image of a quaternary linear code is a linear code. On the other hand, in [2] the author introduces a generalization of the Gray map and extends the concept of $\mathbb{Z}_4$-linear codes to $\mathbb{Z}_{2^k}$-linear codes in a natural way. He also studies the conditions under which the binary codes will be $\mathbb{Z}_{2^k}$-linear; in particular, a characterization of $\mathbb{Z}_8$-linear codes is given.

In this paper, for $k \geq 1$, an isometry between codes over $\mathbb{Z}_{2^{k+1}}$ and codes over $\mathbb{Z}_4$ is introduced and used together with the usual Gray map to give a generalization of the latter mapping. This generalization of the Gray map is equivalent to the one given in [2], but with this reexpression in terms of the isometry from $\mathbb{Z}_{2^{k+1}}^n$ to $\mathbb{Z}_4^{2^{k-1}n}$, it is now possible to give several characterizations related to codes over $\mathbb{Z}_{2^{k+1}}$, similar to those given in Theorems 4, 5, and 6 of [4] for codes over $\mathbb{Z}_4$.

This paper is organized as follows. In section 2, notation and definitions that will be used throughout are established, and results that will be useful in order to introduce the isometry from $\mathbb{Z}_{2^{k+1}}^n$ into $\mathbb{Z}_4^{2^{k-1}n}$ are also presented. Some properties of this isometry, important for the main results of this work, are given in section 3. In section 4 we extend the concept of $\mathbb{Z}_{2^k}$-linear codes to quaternary codes and give

*Received by the editors November 1, 2001; accepted for publication (in revised form) February 26, 2003; published electronically October 2, 2003.

http://www.siam.org/journals/sidma/17-1/39721.html

[†] Departamento de Matemáticas, Universidad Autónoma Metropolitana–I, 09340 México D.F., México (htr@xanum.uam.mx). This author was partially supported by Red de Criptología, UAM, I-CONACyT, México. This work was partially done while the author was on sabbatical at DGSCA, UNAM.

[‡] Dirección General de Servicios de Cómputo Académico (DGSCA), Universidad Nacional Autónoma de México, 04510 México D.F., México (gerardov@servidor.unam.mx).

necessary and sufficient conditions for a binary or quaternary code to be $\mathbb{Z}_{2^{k+1}}$-linear. In section 5 some examples are presented.

2. Definitions, notation, and preliminary results. Let $\mathbb{F}_2$ be the binary field, and for a positive integer n let $\mathbb{F}_2^n$ be the vector space of all binary vectors of length n. For any positive integer $k \geq 1$ let $\mathbb{Z}_{2^{k+1}}$ be the ring of integers modulo 2^{k+1}, and let $\mathbb{Z}_{2^{k+1}}^n$ be the module of n-tuples with entries in $\mathbb{Z}_{2^{k+1}}$. Addition in $\mathbb{F}_2$ and $\mathbb{F}_2^n$ will be denoted by "$\oplus$," while addition in $\mathbb{Z}_{2^{k+1}}$ and $\mathbb{Z}_{2^{k+1}}^n$ will be denoted by "$+$." Nevertheless, as we will see, the action of "$\oplus$" will also be extended to $\mathbb{Z}_{2^{k+1}}$ and $\mathbb{Z}_{2^{k+1}}^n$.

By a *quaternary code* C of length n [4] we mean a block code, linear or not, over $\mathbb{Z}_4$. Equivalently, by a *binary code* C' of length n we mean a block code, linear or not, over $\mathbb{F}_2$. The definition of the Gray map ϕ from $\mathbb{Z}_4^n$ into $\mathbb{F}_2^{2n}$ is as in [4] and [10]; that is, for all $Z = (z_1, z_2, \dots, z_n) \in \mathbb{Z}_4^n$,

$$\phi(Z) = (r_1(z_1), \dots, r_1(z_n), r_1(z_1) \oplus r_0(z_1), \dots, r_1(z_n) \oplus r_0(z_n)) ,$$

where r_1 and r_0 are two maps from $\mathbb{Z}_4$ into $\mathbb{F}_2$ such that if $z \in \mathbb{Z}_4$, then the 2-adic expansion of z is $z = r_0(z) + 2r_1(z)$. A property of the Gray map (see [10]) is

$$\phi(X + 2Y) = \phi(X) \oplus \phi(2Y) \quad \forall\, X, Y \in \mathbb{Z}_4^n. \tag{1}$$

2.1. Two new operations on $\mathbb{Z}_{2^{k+1}}$. Similar to the Gray map, and for $k \geq 1$, $k+1$ mappings r_i, $i = 0, 1, \dots, k$, from $\mathbb{Z}_{2^{k+1}}$ into $\mathbb{F}_2$ are introduced such that if $a \in \mathbb{Z}_{2^{k+1}}$, the 2-adic expansion of a is $a = r_0(a) + 2r_1(a) + \cdots + 2^k r_k(a)$. Using this expansion of any element in $\mathbb{Z}_{2^{k+1}}$, two new operations, "$\oplus$" and "$\odot$," on $\mathbb{Z}_{2^{k+1}}$ are introduced as follows: if $a, b \in \mathbb{Z}_{2^{k+1}}$, then

$$a \oplus b = \sum_{i=0}^{k} 2^i (r_i(a) \oplus r_i(b)),$$
$$a \odot b = \sum_{i=0}^{k} 2^i (r_i(a) r_i(b)).$$

For $l \in \{1, 2, \dots, k+1\}$ we will use $\zeta_l(a, b)$ to denote the lth carry bit (see, for example, [8]) in the rational sum between the binary numbers[1] $(r_k(a), \dots, r_0(a))$ and $(r_k(b), \dots, r_0(b))$. Thus, $\zeta_1(a, b) = r_0(a) r_0(b)$, and since a recursive relation for $\zeta_l(a, b)$ is

$$\zeta_{l+1}(a, b) = (r_l(a) r_l(b)) \oplus \zeta_l(a, b)(r_l(a) \oplus r_l(b))$$

for $l \in \{1, 2, \dots, k\}$, then a general algebraic expression for $\zeta_l(a, b)$ is

$$\zeta_l(a, b) = \bigoplus_{i=0}^{l-1} (r_i(a) r_i(b)) \prod_{j=i+1}^{l-1} (r_j(a) \oplus r_j(b)) \tag{2}$$

for $l \in \{1, 2, \dots, k+1\}$. Using the carry bits $\zeta_l = \zeta_l(a, b)$, let $c \in \mathbb{Z}_{2^{k+1}}$ be given by

$$c = 2\zeta_1 + 2^2 \zeta_2 + \cdots + 2^k \zeta_k.$$

[1] Here a binary number is considered as a rational number expressed in base 2.

Since $(\zeta_k, \dots, \zeta_1, 0)$ is the carry vector that we obtain from the rational sum between the binary numbers $(r_k(a), \dots, r_0(a))$ and $(r_k(b), \dots, r_0(b))$, then $(a+b) = a \oplus b \oplus c$. An alternative way to express the value $(a+b)$ is through the following proposition.

PROPOSITION 2.1. *For any $k \geq 1$, let a and b be in $\mathbb{Z}_{2^{k+1}}$. Then*

$$a + b = (a \oplus b) + 2(a \odot b).$$

Proof. The proof is derived by induction on k. Clearly, the proposition is true when $k = 1$. Suppose that the relation is valid for $k \geq 1$, and let a and b be in $\mathbb{Z}_{2^{k+2}}$. By the induction hypothesis we have

$$(a - r_0(a) + b - r_0(b))/2 = \sum_{i=1}^{k+1} 2^{i-1}[(r_i(a) \oplus r_i(b)) + 2(r_i(a) r_i(b))] \; .$$

This relation is equivalent to

$$a + b = r_0(a) + r_0(b) + \sum_{i=1}^{k+1} 2^{i}[(r_i(a) \oplus r_i(b)) + 2(r_i(a) r_i(b))] \; .$$

Since $r_0(a) + r_0(b) = (r_0(a) \oplus r_0(b)) + 2(r_0(a) r_0(b))$, the proof comes directly. □

2.2. A mapping construction. For $k \geq 2$, define $\rho_k : \mathbb{Z}_{2^{k+1}} \to \mathbb{F}_2^{k-1}$ as $\rho_k(a) = (r_{k-1}(a), \dots, r_2(a), r_1(a))$. For all $i \in \{0, 1, \dots, 2^{k-1}-1\}$, let $\alpha_i^k \in \mathbb{F}_2^{k-1}$ be the binary expression of i using $k-1$ bits, where the most significant bit is on the left; e.g., if $k = 5$ and $i = 13$, then $\alpha_{13}^5 = (1101)$. By means of ρ_k and α_i^k, the following functions $\varphi_i^k : \mathbb{Z}_{2^{k+1}} \to \mathbb{Z}_4$ are introduced:

$$\varphi_i^k(a) = 2[r_k(a) \oplus (\rho_k(a) \cdot \alpha_i^k)] + r_0(a) \tag{3}$$

for all $i = 0, 1, \dots, 2^{k-1} - 1$, where "$\cdot$" denotes the usual dot product in $\mathbb{F}_2^{k-1}$. The action of the functions φ_i^k are extended to $\mathbb{Z}_{2^{k+1}}^n$ as follows: if $A = (a_1, a_2, \dots, a_n) \in \mathbb{Z}_{2^{k+1}}^n$, then $\varphi_i^k(A) = (\varphi_i^k(a_1), \varphi_i^k(a_2), \dots, \varphi_i^k(a_n))$. Thus, the map $\varphi^k : \mathbb{Z}_{2^{k+1}}^n \to \mathbb{Z}_4^{2^{k-1}n}$ is introduced:

$$\varphi^k(A) = (\varphi_0^k(A), \varphi_1^k(A), \dots, \varphi_{2^{k-1}-1}^k(A)) \qquad \forall\, A \in \mathbb{Z}_{2^{k+1}}^n.$$

For completeness, $\varphi^1 : \mathbb{Z}_4^n \to \mathbb{Z}_4^n$ is defined as the identity map, $\varphi^1(A) = A$. Using the map φ^k, an equivalent definition of the *generalized Gray map* $\Phi : \mathbb{Z}_{2^{k+1}}^n \to \mathbb{F}_2^{2^k n}$, as introduced in [2], can be given:

$$\Phi(A) = \phi(\varphi^k(A)).$$

A more general Gray map over chain rings has been introduced recently (cf. [3]).

Note that the mappings Φ and φ^k are both injective. The following properties of these mappings are obvious from the definitions.

PROPOSITION 2.2. *Let $a \in \mathbb{Z}_{2^{k+1}}$, and let $\bar{x} \in \mathbb{F}_2^n$; then*

$$\begin{aligned} \varphi^k(a\bar{x}) &= \varphi^k(a) \otimes \bar{x}, \\ \Phi(a\bar{x}) &= \Phi(a) \otimes \bar{x}, \end{aligned} \tag{4}$$

where "$\otimes$" is the Kronecker product (for example, see [7, Chap. 14, p. 421]*).*

We extend the application of the operations "+," "$\oplus$," and "$\odot$" on $\mathbb{Z}_{2^{k+1}}$ to $\mathbb{Z}_{2^{k+1}}^n$ in the natural way; that is, if $A = (a_1, \dots, a_n), B = (b_1, \dots, b_n) \in \mathbb{Z}_{2^{k+1}}^n$, then we define $A \star B = (a_1 \star b_1, \dots, a_1 \star b_n)$, where the operation $\star$ is any one of these operations. Consequently, the equations $a+b = a \oplus b \oplus c$ and $a+b = (a\oplus b) + 2(a \odot b)$ are extended as

$$\begin{aligned} A + B &= A \oplus B \oplus \left(\sum_{l=1}^{k} 2^l \bar{\zeta}_l \right), \\ A + B &= (A \oplus B) + 2(A \odot B) \end{aligned} \tag{5}$$

for all $A, B \in \mathbb{Z}_{2^{k+1}}^n$, where $\bar{\zeta}_l = (\zeta_l(a_1, b_1), \dots, \zeta_l(a_n, b_n))$.

On the other hand, with the introduction of the operation "$\odot$," we recall another important property of the Gray map.

Proposition 2.3. *Let $X, Y \in \mathbb{Z}_4^n$; then*

$$\begin{aligned} \phi(X + Y + 2(X \odot Y)) &= \phi(X + Y) \oplus \phi(2(X \odot Y)) \\ &= \phi(X) \oplus \phi(Y). \end{aligned} \tag{6}$$

Proof. The proof follows from (1) and the proof of Theorem 5 in [4]. □

2.3. The weight functions. The *Lee weight*, wt_L, of $0, 1, 2, 3 \in \mathbb{Z}_4$ is $0, 1, 2, 1$, respectively, and the Lee weight $wt_L(A)$ of $A \in \mathbb{Z}_4^n$ is the rational sum of the Lee weights of its components. The *Lee distance*, d_L, is defined as $d_L(A, B) = wt_L(A - B)$ for all $A, B \in \mathbb{Z}_4^n$. For $k \geq 1$, the *homogeneous weight*, wt_{hom} on $\mathbb{Z}_{2^{k+1}}$ is defined as [3, 5]

$$wt_{\text{hom}}(a) = \begin{cases} 0 & \text{if } a = 0, \\ 2^k & \text{if } a = 2^k \\ 2^{k-1} & \text{otherwise.} \end{cases} \qquad \forall\, a \in \mathbb{Z}_{2^{k+1}},$$

Again, for $A \in \mathbb{Z}_{2^{k+1}}^n$, the value $wt_{\text{hom}}(A)$ is taken as the rational sum of the homogeneous weights of its components, and the *homogeneous distance*, δ_{hom}, is given by $\delta_{\text{hom}}(A, B) = wt_{\text{hom}}(A - B)$ for all $A, B \in \mathbb{Z}_{2^{k+1}}^n$.

2.4. The new map is an isometry. For $n = 1$, $k \geq 2$, and for $a \in \mathbb{Z}_{2^{k+1}}$, the function $f(i) = r_k(a) \oplus (\rho_k(a) \cdot \alpha_i^k)$ in (3) is an affine Boolean function in $k-1$ variables. Hence, the vector $\varphi^k(a) = (\varphi_0^k(a), \varphi_1^k(a), \dots, \varphi_{2^{k-1}-1}^k(a))$ takes one of the following forms:

- the null vector if $a = 0$;
- all of its entries are equal to 2 if $a = 2^k$;
- half of its entries are equal to 2 and the other half are equal to 0 if a is an even number different from 0 and 2^k;
- all of its entries have the value 1 or 3 if a is odd.

The conclusion is, in any case, that $wt_{\text{hom}}(a) = wt_L(\varphi^k(a))$. Thus, we have the following proposition.

Proposition 2.4. *The map φ^k is an isometry from $(\mathbb{Z}_{2^{k+1}}^n, \delta_{\text{hom}})$ into $(\mathbb{Z}_4^{2^{k-1}n}, d_L)$.*

In [4] the authors take the usual binary Hamming distance d_H on $\mathbb{F}_2^{2n}$ in order to prove that the Gray map ϕ, from $\mathbb{Z}_4^n$ into $\mathbb{F}_2^{2n}$, is an isometry. Thus as a consequence of the previous proposition we have, as in [2], the following corollary.

COROLLARY 2.5. *The generalized Gray map Φ is an isometry from $(\mathbb{Z}^n_{2^{k+1}},\, \delta_{\text{hom}})$ into $(\mathbb{F}_2^{2^k n},\, d_H)$.*

Therefore, since φ^k is an isometry between codes over the rings $\mathbb{Z}_{2^{k+1}}$ and $\mathbb{Z}_4$, it will be called the *modular reduction isometry of order k.*

3. Some properties of the modular reduction isometry. In this section we provide some properties and definitions related to the modular reduction isometry that will be important in order to obtain the main results.

PROPOSITION 3.1. *Let $A, B \in \mathbb{Z}^n_{2^{k+1}}$. If $\bar{r}_0(A)$ is a binary vector of length n that has a one in its ith entry if and only if the ith entry of A is odd for $i = 1, \ldots, n$, then*

$$\begin{aligned} \varphi^k(2^k A + B) &= \varphi^k(2^k A) + \varphi^k(B), \\ -\varphi^k(A) &= \varphi^k(2^k A + A), \\ \varphi^k(2^k A) &= 2\varphi^k(A) = \bar{1} \otimes 2\bar{r}_0(A). \end{aligned} \tag{7}$$

Proof. Without loss of generality we may assume $n = 1$. Thus, from (3) we have $\varphi^k_i(2^k A + B) = 2[r_0(A) \oplus r_k(B) \oplus (\rho_k(B) \cdot \alpha^k_i)] + r_0(B) = 2r_0(A) + 2[r_k(B) \oplus (\rho_k(B) \cdot \alpha^k_i)] + r_0(B) = \varphi^k_i(2^k A) + \varphi^k_i(B)$ for all $i = 0, 1, \ldots, 2^{k-1} - 1$. On the other hand, $-\varphi^k_i(A) = 2[r_k(A) \oplus (\rho_k(A) \cdot \alpha^k_i)] + 3r_0(A) = 2[r_0(A) \oplus r_k(A) \oplus (\rho_k(A) \cdot \alpha^k_i)] + r_0(A) = \varphi^k_i(2^k A + A)$. The final part is straightforward. $\square$

PROPOSITION 3.2. *Let $A \in \mathbb{Z}^n_{2^{k+1}} \subset \mathbb{Z}^n_{2^{k+2}}$. Identifying $\mathbb{Z}_4^{2^k n}$ with $\mathbb{Z}_4^{2^{k-1} n} \times \mathbb{Z}_4^{2^{k-1} n}$, we have*

$$\varphi^{k+1}(A) = (\varphi^k(A \triangleleft 2^k \bar{1}), \varphi^k(A)), \tag{8}$$

where $X \triangleleft Y = (x_1 \ Mod \ y_1, \ldots, x_n \ Mod \ y_n)$ for all $X = (x_1, \ldots, x_n), Y = (y_1, \ldots, y_n) \in \mathbb{Z}^n$.

Proof. Again we may assume $n = 1$. Clearly, the proposition is true when $k = 1$. If $k \geq 2$, the proposition is also valid because

$$\rho_{k+1}(A) \cdot \alpha^{k+1}_i = \begin{cases} \rho_k(A) \cdot \alpha^k_i & \text{if } 0 \leq i \leq 2^{k-1} - 1, \\ r_k(A) \oplus \rho_k(A) \cdot \alpha^k_i & \text{if } 2^{k-1} \leq i \leq 2^k - 1\,. \end{cases} \qquad \square$$

PROPOSITION 3.3. *Let $A = (a_1, \ldots, a_n)$ and $B = (b_1, \ldots, b_n)$ be elements of $\mathbb{Z}^n_{2^{k+1}}$; then*

$$\varphi^k(2^k \bar{\zeta}_1) = 2(\varphi^k(A) \odot \varphi^k(B)), \tag{9}$$

where $\bar{\zeta}_1 = (\zeta_1(a_1, b_1), \ldots, \zeta_1(a_n, b_n))$.

Proof. First note that $\varphi^k(2^k \bar{\zeta}_1) = 2\varphi^k(\bar{\zeta}_1)$. The values $\varphi^k_i(\zeta_1(a_j, b_j))$ and $\varphi^k_i(a_j) \odot \varphi^k_i(b_j) \odot 1$ for $i = 0, \ldots, 2^{k-1} - 1$, $j = 1, \ldots, n$ are binary; that is, such values take only 0 or 1 depending on the input entries a_j and b_j. Thus, $\varphi^k_i(\zeta_1(a_j, b_j)) = 1 \;\Leftrightarrow\; a_j$ and b_j are odd numbers $\Leftrightarrow\; \varphi^k_i(a_j)$ and $\varphi^k_i(b_j)$ are also odd numbers $\Leftrightarrow\; \varphi^k_i(a_j) \odot \varphi^k_i(b_j) \odot 1 = 1$, and the claim is proved. $\square$

PROPOSITION 3.4. *The subset $\varphi^k(\mathbb{Z}_{2^{k+1}}) = \{\varphi^k(a) \mid a \in \mathbb{Z}_{2^{k+1}}\}$ of $\mathbb{Z}_4^{2^{k-1}}$ is a $\mathbb{Z}_4$-submodule generated by the set $\{\varphi^k(a) \mid a = 2^j, j = 0, 1, \ldots, k-1\}$.*

Proof. Let $\bar{u} \in \{\varphi^k(a) \mid a \in \mathbb{Z}_{2^{k+1}}\}$; then $a \in \mathbb{Z}_{2^{k+1}}$ exists such that $\varphi^k(a) = \bar{u}$. By (3), $\varphi^k_i(a) = 2r_k(a) + r_0(a) + 2(\rho_k(a) \cdot \alpha^k_i)$, where addition $+$ is taken on $\mathbb{Z}_4$. But for all $i = 0, 1, \ldots, 2^{k-1} - 1$, the functions φ^k_i are such that

$$\varphi^k_i(1) = 1, \quad \varphi^k_i(2^k) = 2, \quad \text{and}$$

$$\varphi^k_i(2^j) = \begin{cases} 2 & \text{if } \alpha^k_{2^j} \cdot \alpha^k_i \neq 0, \\ 0 & \text{otherwise} \end{cases} \quad \text{with } 0 < j < k.$$

Then the value $\varphi_i^k(a)$ can be rewritten as $\varphi_i^k(a) = (2r_k(a) + r_0(a))\varphi_i^k(1) + \sum_{j=1}^{k-1} r_j(a)\varphi_i^k(2^j)$; hence $\varphi^k(a) = (2r_k(a) + r_0(a))\varphi^k(1) + \sum_{j=1}^{k-1} r_j(a)\varphi^k(2^j)$. □

The following definition takes the previous proposition into consideration.

DEFINITION 3.5. *A subset* C *of* $\mathbb{Z}_4^{2^{k-1}n}$ *is* an isometric image of modular reduction of order k, $IIMR_k$, *if for all* $\bar{x} = (x_1, \dots, x_{2^{k-1}n}) \in C$ *and for all* $i \in \{1, \dots, n\}$ *the tuple* $\hat{x} = (x_i, x_{i+n}, \dots, x_{i+(2^{k-1}-1)n}) \in \mathbb{Z}_4^{2^{k-1}}$ *is such that* $\hat{x} \in \varphi^k(\mathbb{Z}_{2^{k+1}})$*, that is, if* $\hat{x}$ *can be expressed as a linear combination of the set* $\{\varphi^k(a) \mid a = 2^j, j = 0, 1, \dots, k-1\}$.

In the case where $k = 2$ we have $\bar{x} = (x_1, \dots, x_{2n}) \in C \subset \mathbb{Z}_4^{2n}$, and if $\bar{u} = \phi(\bar{x}) = (\bar{u}_1, \bar{u}_2, \bar{u}_3, \bar{u}_4) \in \mathbb{F}_2^{4n}$ with $\bar{u}_i \in \mathbb{F}_2^n$ for $i = 1, \dots, 4$, then $\hat{x} = (x_i, x_{i+n}) \in \langle(1,1),(0,2)\rangle$ ⇔ the numbers x_i and x_{i+n} are both even or both odd ⇔

$$\bar{u}_1 \oplus \bar{u}_2 \oplus \bar{u}_3 \oplus \bar{u}_4 = \bar{0}, \tag{10}$$

where $\bar{0}$ is the all-zero word of length n. Thus (10) gives a special characterization of the case when a subset $C \subset \mathbb{Z}_4^{2n}$ is an $IIMR_2$.

4. $\mathbb{Z}_{2^{k+1}}$-linear codes.

We begin our discussion about $\mathbb{Z}_{2^{k+1}}$-linear codes with the following definition and propositions.

DEFINITION 4.1. *We say that a quaternary code* C *of length* $2^{k-1}n$ *is* $\mathbb{Z}_{2^{k+1}}$-linear *if its coordinates can be arranged so that it is the image under* φ^k*, the modular reduction isometry of order* k*, of a linear code* $\mathcal{C}$ *of length* n *over* $\mathbb{Z}_{2^{k+1}}$*. Similarly, we say that a binary code* C' *of length* $2^k n$ *is* $\mathbb{Z}_{2^{k+1}}$-linear *if its coordinates can be arranged so that it is the image under* Φ*, the generalized Gray map, of a linear code* $\mathcal{C}$ *of length* n *over* $\mathbb{Z}_{2^{k+1}}$.

4.1. Computation of $\varphi^k(A + B)$ and $\Phi(A + B)$.

PROPOSITION 4.2. *Let* $A = (a_1, \dots, a_n)$ *and* $B = (b_1, \dots, b_n)$ *be elements of* $\mathbb{Z}_{2^{k+1}}^n$*; then*

$$\varphi^k(A \oplus B) = \varphi^k(A) + \varphi^k(B) + 2\varphi^k(A \odot B)\ . \tag{11}$$

Proof. First observe that $2^k(A \odot B) = 2^k\bar{\zeta}_1$; hence $2\varphi^k(A \odot B) = \varphi^k(2^k\bar{\zeta}_1)$. The rest of the proof is derived by induction over k, and since $A \oplus B = A + B + 2\bar{\zeta}_1$, when $k = 1$, then (11) is true for this case. Suppose that (11) is true for $k \geq 1$, and for $\bar{r}_{k+1}, \bar{r}'_{k+1} \in \mathbb{F}_2^n$ and $A, B \in \mathbb{Z}_{2^{k+1}}^n$, let $A', B' \in \mathbb{Z}_{2^{k+2}}^n$ be given by $A' = (a'_1, \dots, a'_n) = 2^{k+1}\bar{r}_{k+1} + A$ and $B' = (b'_1, \dots, b'_n) = 2^{k+1}\bar{r}'_{k+1} + B$. Note that when $k \geq 2$, $\zeta_1(a'_i, b'_i) = \zeta_1(a_i, b_i)$ for $i = 1, \dots, n$. Now, it follows from (7) and (8) that

$$\varphi^{k+1}(A') = \varphi^{k+1}(2^{k+1}\bar{r}_{k+1}) + (\varphi^k(A \triangleleft 2^k\bar{1}), \varphi^k(A)), \tag{12}$$

and equivalently for $\varphi^{k+1}(B')$. On the other hand, for $\varphi^{k+1}(A' \oplus B')$ we obtain

$$\begin{aligned} \varphi^{k+1}(A' \oplus B') &= \varphi^{k+1}(2^{k+1}(\bar{r}_{k+1} \oplus \bar{r}'_{k+1}) + A \oplus B), \\ &= \varphi^{k+1}(2^{k+1}(\bar{r}_{k+1} \oplus \bar{r}'_{k+1})) \\ &\quad + (\varphi^k((A \oplus B) \triangleleft 2^k\bar{1}), \varphi^k(A \oplus B)). \end{aligned} \tag{13}$$

Since $\varphi^{k+1}(2^{k+1}\bar{\zeta}_1) = (\varphi^k(2^k\bar{\zeta}_1), \varphi^k(2^k\bar{\zeta}_1))$, by the induction hypothesis it is readily seen that $(\varphi^k((A \oplus B) \triangleleft 2^k\bar{1}), \varphi^k(A \oplus B))$ is equal to

$$\varphi^{k+1}(2^{k+1}\bar{\zeta}_1) + (\varphi^k(A \triangleleft 2^k\bar{1}) + \varphi^k(B \triangleleft 2^k\bar{1}), \varphi^k(A) + \varphi^k(B)). \tag{14}$$

Considering (12) and (14) and comparing them with (13), the proof of the claim follows. ∎

Proposition 4.3. *Using the notation of the previous proposition, we have*

$$\varphi^k(A+B) = \varphi^k(A \oplus B) + \varphi^k((A+B) \oplus A \oplus B)\ .$$

Proof. Since all of the entries of the vector $(A+B) \oplus A \oplus B$ are even numbers, by the previous proposition we conclude that $\varphi^k(A+B) = \varphi^k(A \oplus B \oplus (A+B) \oplus A \oplus B) = \varphi^k(A \oplus B) + \varphi^k((A+B) \oplus A \oplus B)$. ∎

Proposition 4.4. *With notation as in Proposition* 4.2, *we have*

$$\Phi(A+B) = \Phi(A) \oplus \Phi(B) \oplus \Phi((A+B) \oplus A \oplus B).$$

Proof. $\Phi(A+B) = \phi(\varphi^k(A) + \varphi^k(B) + \varphi^k((A+B) \oplus A \oplus B + 2^k\bar{\zeta}_1))$. Now, observe that $\varphi^k((A+B) \oplus A \oplus B + 2^k\bar{\zeta}_1)$ is a vector in $\mathbb{Z}_4^{2^{k-1}n}$ of the form $2X$, with $X \in \mathbb{F}_2^{2^{k-1}n}$. From (1), we have $\Phi(A+B) = \phi(\varphi^k(A) + \varphi^k(B)) \oplus \phi(\varphi^k(2^k\bar{\zeta}_1)) \oplus \Phi((A+B) \oplus A \oplus B)$, but from (9) we know that $\varphi^k(2^k\bar{\zeta}_1) = 2(\varphi^k(A) \odot \varphi^k(B))$. Since $\phi(\varphi^k(A) + \varphi^k(B)) \oplus \phi(2(\varphi^k(A) \odot \varphi^k(B))) = \Phi(A) \oplus \Phi(B)$, (see (6)), the claim follows. ∎

4.2. A characterization of $\mathbb{Z}_{2^{k+1}}$-linear codes. A characterization of the $\mathbb{Z}_{2^{k+1}}$-linear codes, similar to the one given in [4] for $\mathbb{Z}_4$-linear codes, is as follows.

Theorem 4.5. *A quaternary code C of length $2^{k-1}n$ is $\mathbb{Z}_{2^{k+1}}$-linear if and only if its coordinates can be arranged so that*

1. *C is an $IIMR_k$;*
2. *for all $\bar{x}, \bar{y} \in C$ with $\bar{x} = \varphi^k(A)$, $\bar{y} = \varphi^k(B)$ for some $A, B \in \mathbb{Z}_{2^{k+1}}^n$; then*

$$\bar{x} + \bar{y} + \varphi^k((A+B) \oplus A \oplus B + 2^k(A \odot B)) \in C.$$

Proof. This is an immediate consequence of Definition 3.5 and Propositions 4.2 and 4.3. ∎

In the case of binary codes we have the following theorem.

Theorem 4.6. *A binary code C' of length $2^k n$ is $\mathbb{Z}_{2^{k+1}}$-linear if and only if its coordinates can be arranged so that the quaternary code $C = \phi^{-1}(C')$ satisfies condition* 1 *of the previous theorem and if for all $\bar{u}, \bar{v} \in C'$ with $\bar{u} = \Phi(A)$, $\bar{v} = \Phi(B)$ for some $A, B \in \mathbb{Z}_{2^{k+1}}^n$, then*

$$\bar{u} \oplus \bar{v} \oplus \Phi((A+B) \oplus A \oplus B) \in C'.$$

Proof. This is an immediate consequence of Definition 3.5 and Proposition 4.4. ∎

By means of relations (1), (4), and (5), $\Phi((A+B) \oplus A \oplus B)$ can be written as

$$\Phi((A+B) \oplus A \oplus B) = \bigoplus_{l=1}^{k} \Phi(2^l) \otimes \bar{\zeta}_l.$$

If $k = 1$, suppose that such a code C' satisfies the conditions in Theorem 4.6. If $\bar{u} = (\bar{u}_1, \bar{u}_2), \bar{v} = (\bar{v}_1, \bar{v}_2) \in C'$ with $\bar{u}_j = (u_{j,1,\dots,u_{j,n}}), \bar{v}_j = (v_{j,1,\dots,v_{j,n}}) \in \mathbb{F}_2^n$ for $j = 1, 2$, then $A = (a_1, \dots, a_n), B = (b_1, \dots, b_n) \in \mathbb{Z}_4^n$ exists such that $\Phi(A) = \bar{u}$ and $\Phi(B) = \bar{v}$. Under these circumstances, we have

$$\begin{aligned} a_i &= 2u_{1,i} + (u_{1,i} \oplus u_{2,i}), \\ b_i &= 2v_{1,i} + (v_{1,i} \oplus v_{2,i}) \end{aligned}$$

for $i \in \{1, \ldots, n\}$. Considering relation (2), $\bar{\zeta}_1 = (\bar{u}_1 \oplus \bar{u}_2) \odot (\bar{v}_1 \oplus \bar{v}_2)$, and since $\Phi(2) = (1,1)$, then a binary code C' of length $2n$ is $\mathbb{Z}_4$-linear if and only if its coordinates can be arranged so that any pair of vectors $\bar{u}, \bar{v} \in C'$ satisfies

$$\bar{u} \oplus \bar{v} \oplus (\bar{\zeta}_1, \bar{\zeta}_1) \in C'.$$

Observe that this condition is the same as that provided in the characterization of $\mathbb{Z}_4$-linear codes given in [4, Thm. 4].

On the other hand, in the case where $k = 2$ suppose again that such a code C' satisfies the conditions in Theorem 4.6. Then if $\bar{u} = (\bar{u}_1, \ldots, \bar{u}_4), \bar{v} = (\bar{v}_1, \ldots, \bar{v}_4) \in C'$ with $\bar{u}_j = (u_{j,1}, \ldots, u_{j,n}), \bar{v}_j = (v_{j,1}, \ldots, v_{j,n}) \in \mathbb{F}_2^n$ for $j = 1, \ldots, 4$, then $A = (a_1, \ldots, a_n), B = (b_1, \ldots, b_n) \in \mathbb{Z}_8^n$ exists such that $\Phi(A) = \bar{u}$ and $\Phi(B) = \bar{v}$. Under these circumstances, we have

$$\begin{aligned} a_i &= 4u_{1,i} + 2(u_{1,i} \oplus u_{2,i}) + (u_{1,i} \oplus u_{3,i}), \\ b_i &= 4v_{1,i} + 2(v_{1,i} \oplus v_{2,i}) + (v_{1,i} \oplus v_{3,i}) \end{aligned}$$

for $i \in \{1, \ldots, n\}$. Again from (2) we have

$$\begin{aligned} \bar{\zeta}_1 &= (\bar{u}_1 \oplus \bar{u}_3) \odot (\bar{v}_1 \oplus \bar{v}_3), \\ \bar{\zeta}_2 &= ((\bar{u}_1 \oplus \bar{u}_2) \odot (\bar{v}_1 \oplus \bar{v}_2)) \oplus ((\bar{u}_1 \oplus \bar{u}_3) \\ &\qquad \odot (\bar{v}_1 \oplus \bar{v}_3) \odot (\bar{u}_1 \oplus \bar{u}_2 \oplus \bar{v}_1 \oplus \bar{v}_2)). \end{aligned}$$

Since $\Phi(2) = (0,1,0,1)$ and $\Phi(4) = (1,1,1,1)$, we conclude that a binary code C' of length $4n$ is $\mathbb{Z}_8$-linear if and only if its coordinates can be arranged so that any pair of vectors $\bar{u}, \bar{v} \in C'$ satisfies (10) and

$$\bar{u} \oplus \bar{v} \oplus (\bar{0}, \bar{\zeta}_1, \bar{0}, \bar{\zeta}_1) \oplus (\bar{\zeta}_2, \bar{\zeta}_2, \bar{\zeta}_2, \bar{\zeta}_2) \in C'.$$

These last conditions are equivalent to those in the characterization of $\mathbb{Z}_8$-linear codes given in [2, Prop. 4].[2]

As a consequence of the previous theorems a generalization of Theorem 6 of [4] for quaternary and binary linear codes can be given.

Corollary 4.7. *A quaternary linear code C of length $2^{k-1}n$ is $\mathbb{Z}_{2^{k+1}}$-linear if and only if its coordinates can be arranged so that the code C satisfies condition* 1 *of Theorem* 4.5 *and if for all $\bar{x}, \bar{y} \in C$ with $\bar{x} = \varphi^k(A)$, $\bar{y} = \varphi^k(B)$ for some $A, B \in \mathbb{Z}_{2^{k+1}}^n$, then*

$$\varphi^k((A + B) \oplus A \oplus B + 2^k(A \odot B)) \in C.$$

Corollary 4.8. *A binary linear code C' of length $2^k n$ is $\mathbb{Z}_{2^{k+1}}$-linear if and only if its coordinates can be arranged so that the quaternary code $C = \phi^{-1}(C')$ satisfies condition* 1 *of Theorem* 4.5 *and if for all $\bar{u}, \bar{v} \in C'$ with $\bar{u} = \Phi(A)$, $\bar{v} = \Phi(B)$ for some $A, B \in \mathbb{Z}_{2^{k+1}}^n$, then*

$$\Phi((A + B) \oplus A \oplus B) \in C'.$$

[2] In fact, in [2] a different ordering in the definition of the generalized Gray map was used. Also, there is a slight misprint in this paper. For this reason, the conditions given here are somewhat different from those in [2].

4.3. Linear gray map images of codes over $\mathbb{Z}_{2^{k+1}}$. If $\mathcal{C}$ is a code over $\mathbb{Z}_{2^{k+1}}$, linear or not, what conditions must be satisfied in order for its image to be linear under the generalized Gray map or under the modular reduction isometry of order k? The following results give an answer to this question.

THEOREM 4.9. *Let $\mathcal{C}$ be a nonempty code, linear or not, over $\mathbb{Z}_{2^{k+1}}$. Then the quaternary code $C = \varphi^k(\mathcal{C})$ is linear if and only if*

$$\forall\ A, B \in \mathcal{C},\ \text{then}\ A \oplus B \oplus 2^k(A \odot B) \in \mathcal{C}. \tag{15}$$

Proof. This is an easy consequence of Propositions 3.1 and 4.2 and the fact that when (15) is true, we have

$$\{\lambda_1 A \oplus \lambda_2 B \oplus 2^k \bar{\zeta}_1(r_0(\lambda_1)A, r_0(\lambda_2)B) \mid \\ \lambda_1, \lambda_2 \in \{0, 1, 2^k, 2^k+1\}\} \subseteq \mathcal{C}. \qquad \square$$

In [4] the authors give a characterization (Theorem 5) on the linearity of the Gray map image of a quaternary linear code, given in the following theorem.

THEOREM 4.10. *The binary image $C' = \phi(C)$ of a quaternary linear code C is linear if and only if*

$$\forall\ \bar{x}, \bar{y} \in C,\ \text{then}\ 2(\bar{x} \odot \bar{y}) \in C.$$

As a consequence of the last two theorems and Propositions 3.3 and 4.2, we obtain the following theorem.

THEOREM 4.11. *Let $\mathcal{C}$ be a nonempty code, linear or not, over $\mathbb{Z}_{2^{k+1}}$, and assume that it is closed under the operation "$\oplus$"; that is, $A \oplus B \in \mathcal{C}$ for all $A, B \in \mathcal{C}$. If such a code satisfies*

$$\forall\ A, B \in \mathcal{C},\ \text{then}\ 2^k(A \odot B) \in \mathcal{C},$$

then the quaternary image $\varphi^k(\mathcal{C})$ and the binary image $\Phi(\mathcal{C})$ are both linear.

Observe that condition (15) does not imply that $2^k\bar{\zeta}_1 \in \mathcal{C}$. In fact, codes over $\mathbb{Z}_{2^{k+1}}$ that satisfy (15) but whose images under the generalized Gray map are not linear do exist. An example of this is when $k = 1$ and $n = 3$ with the quaternary linear code given by $\mathcal{C} = \langle(1,0,1),(1,1,0)\rangle$.

4.4. Linear gray map images of linear codes over $\mathbb{Z}_{2^{k+1}}$. Now we will give a generalization of Theorem 4.10 for the generalized Gray map Φ and for the modular reduction isometry φ^k.

THEOREM 4.12. *The quaternary image $\varphi^k(\mathcal{C})$ of a linear code $\mathcal{C}$ over $\mathbb{Z}_{2^{k+1}}$ is linear if and only if*

$$\forall\ A, B \in \mathcal{C},\ \text{then}\ (2^k - 2)(A \odot B) \in \mathcal{C}.$$

Proof. We know that $A + B = (A \oplus B) + 2(A \odot B)$; hence $A + B = (A \oplus B \oplus 2^k(A \odot B)) + (2^k + 2)(A \odot B)$. On the other hand, $-(2^k + 2) = (2^k - 2)$; thus $A + B + (2^k - 2)(A \odot B) = A \oplus B \oplus 2^k(A \odot B)$, and the result comes from Theorem 4.9. $\square$

THEOREM 4.13. *Let $\mathcal{C}$ be a linear code over $\mathbb{Z}_{2^{k+1}}$; then for $k > 1$ the following statements are equivalent.*

1. *$\varphi^k(\mathcal{C})$ is linear.*
2. *$\Phi(\mathcal{C})$ is linear.*
3. *$2(A \odot B) \in \mathcal{C}$ for all $A, B \in \mathcal{C}$.*

Proof. Equivalence of statements 2 and 3: From relation (5) $A+B-2(A\odot B) = (A\oplus B)$; thus from Proposition 4.2, we have $\varphi^k(A+B-2(A\odot B)) = \varphi^k(A)+\varphi^k(B)+\varphi^k(2^k\bar{\zeta}_1)$. Using (1), (6), and (9), it follows that

$$\Phi(A+B-2(A\odot B)) = \Phi(A)\oplus\Phi(B).$$

Equivalence of statements 1 and 3: Since $k>1$, then $(2^k-2)(A\odot B)\in\mathcal{C}$ if and only if $2(A\odot B)\in\mathcal{C}$, and this equivalence comes from Theorem 4.12. □

5. Examples. In the following examples it will be useful to represent elements of $\mathbb{Z}_{2^{k+1}}^n$ as polynomials in the ring $\mathbb{Z}_{2^{k+1}}[x]/(x^n-1)$. This is achieved via the polynomial representation map given by

$$(a_1,a_2,\ldots,a_n)\mapsto a_1+a_2x+\cdots+a_nx^{n-1}\ .$$

The operation "$\odot$" introduced in section 2 can be extended to the quotient ring $\mathbb{Z}_{2^{k+1}}[x]/(x^n-1)$ as follows: if $A(x)=\sum_{j=1}^n a_jx^{j-1}$ and $B(x)=\sum_{j=1}^n b_jx^{j-1}$ are elements of $\mathbb{Z}_{2^{k+1}}[x]/(x^n-1)$, then

$$A(x)\odot B(x) = \sum_{j=1}^n (a_j\odot b_j)x^{j-1}\ .$$

Example 1. The binary Hamming code of order 3 is a linear cyclic code generated by the polynomial x^3+x+1. Hensel-lifting this polynomial to $\mathbb{Z}_8[x]$ results in the polynomial $G(x)=x^3+6x^2+5x-1$, which generates a linear cyclic code $\mathcal{C}$ over $\mathbb{Z}_8$ (for linear cyclic codes over weak structures see, for example, [1, 6]). Since this polynomial does not divide $2(G(x)\odot 2G(x))=4$, then by Theorem 4.13 the images $\varphi^2(\mathcal{C})$ and $\Phi(\mathcal{C})$ are nonlinear.

Example 2. Over $\mathbb{F}_2[x]$ we have $x^7-1=(x-1)(x^3+x+1)(x^3+x^2+1)$. By Hensel-lifting to $\mathbb{Z}_8$ we obtain $x^7-1=(x-1)(x^3+6x^2+5x-1)(x^3+3x^2+2x-1)$. Let $\mathcal{C}$ be the linear cyclic code of length 7 over $\mathbb{Z}_8$ generated by the polynomial $G(x)=(x^3+6x^2+5x-1)((x^3+3x^2+2x-1)+4(x-1))$. Since $2((G(x)U_1(x))\odot(G(x)U_2(x)))=G(x)(2(U_1(1)\odot U_2(1)))$ for all $U_1(x),U_2(x)\in\mathbb{Z}_8[x]/(x^7-1)$, then by Theorem 4.13 the images $\varphi^2(\mathcal{C})$ and $\Phi(\mathcal{C})$ are linear.

6. Conclusions. An isometry between the modules $\mathbb{Z}_{2^{k+1}}^n$ and $\mathbb{Z}_4^{n2^{k-1}}$ was introduced which helps to provide several characterizations on $\mathbb{Z}_{2^{k+1}}$-linear codes by means of the generalized Gray map. These results generalize those appearing as Theorems 4, 5, and 6 in [4] for quaternary codes. Also, by means of Theorem 4.13 a family of linear constacyclic codes over $\mathbb{Z}_{2^{k+1}}$ whose images under the generalized Gray map are binary linear quasi-cyclic codes was obtained (see [9] for details).

Acknowledgment. The authors would like to thank C. Carlet for helpful discussions and for pointing out a proof of Proposition 2.1.

REFERENCES

[1] A. R. Calderbank and N. J. A. Sloane, *Modular and p-adic cyclic codes*, Des. Codes Cryptogr., 6 (1995), pp. 21–35.

[2] C. Carlet, *$\mathbb{Z}_{2^k}$-linear codes*, IEEE Trans. Inform. Theory, 44 (1998), pp. 1543–1547.

[3] M. Greferath and S. E. Schmidt, *Gray isometries for finite chain rings and a nonlinear ternary (*$36,3^{12},15$*) code*, IEEE Trans. Inform. Theory, 45 (1999), pp. 2522–2524.

[4] A. R. Hammons, Jr., P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Solé, *The $\mathbb{Z}_4$-linearity of Kerdock, Preparata, Goethals, and related codes*, IEEE Trans. Inform. Theory, 40 (1994), pp. 301–319.

[5] W. Heise, T. Honold, and A. A. Nechaev, *Weighted modules and representations of codes*, in Proceedings of the Sixth International Workshop on Algebraic and Combinatorial Coding Theory, Pskov, Russia, 1998, pp. 123–129.

[6] P. Kanwar and S. R. Lopez-Permounth, *Cyclic codes over the itegers modulo p^m*, Finite Fields Appl., 3 (1997), pp. 334–352.

[7] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.

[8] H. Troy Nagle, Jr., B. D. Carroll, and J. D. Irwin, *An Introduction to Computer Logic*, Prentice-Hall, Englewood Cliffs, N.J., 1975.

[9] H. Tapia-Recillas and G. Vega, *Some constacyclic codes over Z_{2^k} and binary quasicyclic codes*, Discrete Appl. Math., 128 (2003), pp. 305–316.

[10] J. Wolfmann, *Negacyclic and cyclic codes over $\mathbb{Z}_4$*, IEEE Trans. Inform. Theory, 45 (1999), pp. 2527–2532.

SIAM J. Discrete Math.
Vol. 17, No. 1, pp. 114–121

THE COMPLEXITY OF THE EXTENDIBILITY PROBLEM FOR FINITE POSETS*

BENOIT LAROSE† AND LÁSZLÓ ZÁDORI‡

Abstract. For a finite poset P let EXT(P) denote the following decision problem. Given a finite poset Q and a partial map f from Q to P, decide whether f extends to a monotone total map from Q to P.

It is easy to see that EXT(P) is in the complexity class **NP**. In [*SIAM J. Comput.*, 28 (1998), pp. 57–104], Feder and Vardi define the classes of width 1 and of bounded strict width constraint satisfaction problems for finite relational structures. Both classes belong to the broader class of bounded width problems in **P**. We prove that for any finite poset P, if EXT(P) has bounded strict width, then it has width 1. In other words, if a poset admits a near unanimity operation, it also admits a totally symmetric idempotent operation of any arity. In [*Fund. Inform.*, 28 (1996), pp. 165–182], Pratt and Tiuryn proved that SAT(P), a polynomial-time equivalent of EXT(P) is **NP**-complete if P is a crown. We generalize Pratt and Tiuryn's result on crowns by proving that EXT(P), is **NP**-complete for any finite poset P which admits no nontrivial idempotent Malcev condition.

Key words. **NP**-complete, extendibility problem for posets, constraint satisfaction problem, zigzags, totally symmetric idempotent and near unanimity operations

AMS subject classifications. 06A07, 68A20, 03C05

DOI. 10.1137/S0895480101389478

1. Introduction.

We consider the following decision problem: Let P and Q be two finite posets and f a partial map from Q to P. Does f extend to a monotone total map from Q to P?

We get different problems depending on which of P and Q is considered as an input. As easily observed, the problem obtained in each case is contained in complexity class **NP**. The case when the problem takes P, Q, and f as inputs is **NP**-complete, as was proved by Duffus and Goddard in [5]. For a fixed finite poset P let EXT(P), denote the above problem with inputs Q and f. We call EXT(P) the extendibility problem for P. The complexity of a decision problem called SAT(P), a polynomial-time equivalent of EXT(P), was studied by Pratt and Tiuryn in [13]. They proved that if P is a crown, then SAT(P) is **NP**-complete and also gave examples of P when SAT(P) is in **P** .

The *height* of a finite poset P is defined to be the maximum value of k such that there exists a $(k+1)$-element chain in P. A finite poset is *dismantlable* whenever its elements can be listed in such a way that each element, except the last one, has a unique upper or lower cover in the subposet determined by it and its successors. A connected finite poset is *ramified* if it has at least two elements and no element with a unique upper or lower cover.

A *relational set (or structure)* is defined here to be a set equipped with finitary relations. A relational set always is of a certain *type* and it is *finite* if its base set is

*Received by the editors April 20, 2001; accepted for publication (in revised form) May 7, 2003; published electronically October 2, 2003.

http://www.siam.org/journals/sidma/17-1/38947.html

†Department of Mathematics and Statistics, Concordia University, 1455 de Maisonneuve West, Montreal, QC, Canada, H3G 1M8 (larose@mathstat.concordia.ca). The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada.

‡Bolyai Intézet, Aradi vértanúk tere 1, H-6720 Szeged, Hungary (zadori@math.u-szeged.hu). The research of this author was partially supported by OTKA grants T022867, T026243 and by FKFP grant 1259/1997.

finite. A *reduct* of a relational set R is a relational set obtained from R by leaving out some relations. For a finite relational set R we define a decision problem denoted by $\mathrm{CSP}(R)$ that is called the constraint satisfaction problem for R. An instance of $\mathrm{CSP}(R)$ is a finite-type reduct R' of R and a finite relational set X similar to R', and the question is whether there is a morphism from X to R'.

In the case when R is of finite type, each instance I related to R' is considered as an instance related to R by supplying I with additional empty relations, one for each relational symbol missing in the type of R'. So if R is of finite type, we conceive $\mathrm{CSP}(R)$ as a decision problem whose instances are the finite structures similar to R, and the question for an instance I is if there is a homomorphism from I to R. This point of view is followed by Feder and Vardi in [6] and makes only constant time difference in complexity compared with the general definition given for $\mathrm{CSP}(R)$ in the preceding paragraph.

In [6] Feder and Vardi proved that for every relational set R of finite type $\mathrm{CSP}(R)$ is polynomial-time equivalent to the retraction problem for some finite poset P. The latter problem is easily shown to be equivalent to $\mathrm{SAT}(P)$; see [13]. It also turned out in [6] that P might be assumed to be of height at most 2 and ramified. We mention that, in contrast with this result, Pratt and Tiuryn proved in [13] that for every poset P of height 1, $\mathrm{SAT}(P)$ is **NP**-complete if P contains a crown and is in **P** otherwise. It is clear from the above result of Feder and Vardi that statements on the complexity of the extendibility problem for posets might have a bearing on the complexity of CS problems.

In [6] Feder and Vardi defined the notions of width and strict width for CS problems. They showed that CS problems of bounded width have polynomial-time complexity. CS problems of strict width k are certain problems of width k. In [6] it was proved that $\mathrm{CSP}(R)$ has bounded strict width if and only if R admits a near unanimity operation. It was also shown that $\mathrm{CSP}(R)$ has width 1 if and only if there is a certain structure that admits a homomorphism into R. Width 1 problems were characterized by the existence of a set function by Dalmau and Pearson in [3].

In the present paper we shall investigate the complexity of $\mathrm{EXT}(P)$ for a finite poset P. It turns out that there is a strong correlation between the complexity of the extendibility problem and the set of the monotone idempotent operations of the poset.

Algebras whose term operations coincide with the monotone operations of a poset are called *order primal.* Idempotent Malcev conditions satisfied in order primal algebras received much attention in the last decade. Some special operations play an important role in these investigations. Let f be an n-ary operation. We say that f is *symmetric (cyclic)* if it obeys the identity

$$f(x_1, \dots, x_n) = f(x_{\pi(1)}, \dots, x_{\pi(n)})$$

for every permutation (for an n-cycle) π of $\{1, \dots, n\}$. The operation f is called *totally symmetric* if it obeys the identity

$$f(x_1, \dots, x_n) = f(y_1, \dots, y_n)$$

for all sets of variables such that $\{x_1, \dots, x_n\} = \{y_1, \dots, y_n\}$. An n-ary operation f is *idempotent* if it obeys the identity

$$f(x, x, \dots, x) = x.$$

Let $n \geq 3$. An n-ary operation f is called a *near unanimity operation* if the following n identities hold:

$$f(x,\dots,x,\underset{\smile}{\underset{i}{y}},x,\dots,x)=x, \qquad i\in\{1,\dots,n\}.$$

The main results on special idempotent term operations in order primal algebras that we rely on are contained in [12] and [16]. In [12] the authors characterized the finite posets admitting a near unanimity operation. In [16] Szabó and Zádori gave a characterization of finite posets admitting a totally symmetric idempotent (TSI) operation of arity equal to the size of the poset. We note that the latter characterization can be derived from results on width 1 in [6] and [3].

In the present paper we investigate the relationship between the above-mentioned complexity and algebraic results. We show that if a poset P admits a near unanimity operation, then it admits a TSI operation of arity $|P|$. This means that bounded strict width implies width 1 for extendibility problems of posets. We give an example showing that a similar implication does not hold for CS problems of relational structures. Further, we prove that $\mathrm{EXT}(P)$ is polynomial-time equivalent to a CS problem. Then by invoking Jeavons's ideas in [9] we generalize Pratt and Tiuryn's **NP**-completeness result on crowns.

2. Results. First we describe the relationship between two classes of extendibility problems for finite posets. Both of these classes lie in **P** and are relativized versions of important classes of CS problems.

We call a pair (Q,f) a *P-colored poset* if Q is a poset and f is a partial map from Q to P. The P-colored poset (Q,f) or the partial map f is called *P-extendible* if f extends to a monotone total map from Q to P. So we might consider $\mathrm{EXT}(P)$ to be the problem whose inputs are just the finite P-colored posets, and the task is to decide the extendibility of a P-colored poset.

Let (Q,f) and (H,g) be two P-colored posets. A monotone map α from Q to H is called a *homomorphism* from (Q,f) to (H,g) if $f=g\alpha$. The colored poset (H,g) is a *homomorphic image* of (Q,f) under the homomorphism α if α is onto. Now, $\mathrm{EXT}(P)$ might be considered to be the problem whose input is a finite P-colored poset (H,g), and the task is to decide if there is a homomorphism from (H,g) to the P-colored poset (P,Id_P), where Id_P is the identity map on P. Since in a CS problem the task also is to decide if there exists a homomorphism to a designated structure from a similar structure, the notions and theorems related to the CS problems in [6] easily transfer to extendibility problems of posets.

Note that there is a subtle difference that prevents us from considering $\mathrm{EXT}(P)$ to be a CS problem, namely that the instances of $\mathrm{EXT}(P)$ are certain poset structures while an instance for a CS problem is allowed to be any relational structure. Later in Proposition 9, we shall see that up to polynomial-time equivalence one can safely drop this restriction. For the time being we insist on this distinction.

Let P be a finite poset. For $\mathrm{EXT}(P)$ the notions of width and strict width are defined in the same manner as those for CS problems in [6] of Feder and Vardi. By transferring the proofs of Theorems 19, 20, 22, and 24 in [6] and Theorem 1 in [3], one gets that $\mathrm{EXT}(P)$ has width 1 if and only if P admits a TSI operation of arity $|P|$ and that $\mathrm{EXT}(P)$ has bounded strict width if and only if P admits a near unanimity operation.

Let H and Q be finite posets. We say that Q *contains* H if the base set of Q contains the base set of H and the order relation of Q contains the order relation of

H. A finite P-colored poset (Q, f) is called a *zigzag* if it is nonextendible but for any H properly contained in Q the P-colored poset $(H, f \restriction_H)$ is extendible. A P-colored poset (Q, f) is a *tree* if the covering graph of Q is a tree. Let (Q, f) and (H, g) be two P-colored posets. For a finite poset P let $I(P)$ be the algebra whose base set is the base set of P and whose basic operations are the monotone idempotent operations of P. A subalgebra of a finite power of $I(P)$ is called an *idempotent P-subalgebra*. Note that the idempotent P-subalgebras inherit an order structure from P. Hence we shall consider them as posets. The notions defined in this paragraph play an important role in characterizations of finite posets that admit a near unanimity operation or a TSI operation of arity equal to the size of the base set of the poset.

Theorem 1 (see [12]). *For a finite connected poset P the following are equivalent:*

(1) *P admits a near unanimity operation.*

(2) *P has finitely many P-zigzags.*

(3) *The idempotent P-subalgebras are dismantlable.*

It is easy to prove (see [16]) that a finite poset P admits a TSI operation of arity $|P|$ if and only if P admits an n-ary TSI operation for all n.

Theorem 2 (see [16]). *A finite poset P admits a TSI operation of arity $|P|$ if and only if every P-zigzag is a homomorphic image of a nonextendible tree.*

Now, we are ready to prove that bounded strict width implies width 1 in the class of the extendibility problems for finite posets. By our earlier note it suffices to prove the following.

Theorem 3. *Every finite poset P which admits a near unanimity operation also admits a TSI operation of arity $|P|$.*

Proof. We refer to the proof of Theorem 4 in [16]. In the proof it is shown that if P has finitely many zigzags and admits a cyclic operation of arbitrary arity, then every P-zigzag is a homomorphic image of a nonextendible tree. Now, we sharpen this result by showing that if P has finitely many zigzags, then P admits a cyclic idempotent operation of any arity. Then by Theorems 1 and 2 we get the claim for finite connected posets. For the case when P is not connected and admits a near unanimity operation observe that the components of P also admit a near unanimity operation. Moreover, the P-zigzags are the zigzags of the components and zigzags whose base posets are fences. Hence, the nonconnected case follows from the connected one.

So for the proof of this theorem let us suppose that P has finitely many P zigzags. Then by Theorem 1 every idempotent P-subalgebra is dismantlable. By a result of Rival in [14], every dismantlable poset has the fixed-point property. We apply this to the P-subalgebra I_n formed by the n-ary monotone idempotent operations of P, $n = 2, 3, \dots$. We define the monotone operation α_n on I_n by

$$\alpha_n : f(x_1, \dots, x_n) \mapsto f(x_2, \dots, x_n, x_1).$$

Since for each n, I_n has the fixed point property, α_n has a fixed point. This fixed point is an n-ary cyclic idempotent operation for each n. □

We remark that there exist finite relational sets of finite type which admit a near unanimity operation but no TSI operation of any arity. For example, let us consider a finite set A with at least two elements and a ternary near unanimity operation m on it. Let R be the relational set whose base set is A and whose relations are all binary relations preserved by m. It is well known in clone theory that, since m is near unanimity, if R admits an operation f, then f is built from m and the projections via composition. But since m is a near unanimity operation, the only binary operations obtained in this way are projections. Hence R admits no binary TSI operation and hence none of any higher arity either.

There are simple examples of posets which admit a TSI operation of every arity but no near unanimity operation, for example, the poset $2+2+1$; see [16]. We do not know if bounded width implies width 1 in the class of the extendibility problems of finite posets.

During the editorial process of this paper Kun and Szabó have published paper [10] in which they describe a polynomial-time algorithm that decides if a finite poset admits a near unanimity operation. This improves the decidability result in [12]. In comparison, no algorithm is known for finite structures of finite type to decide if they admit a near unanimity operation. In the same paper Szabó and Kun actually construct a symmetric idempotent operation of any arity for any poset that admits a near unanimity operation.

Next we shall prove that for every poset which admits no nontrivial idempotent Malcev condition, the extendibility problem is **NP**-complete. It was observed by Corominas in [2] that the only idempotent monotone operations on crowns are the projections. Hence crowns admit no nontrivial idempotent Malcev condition and Pratt and Tiuryn's result on crowns is a consequence of ours. Our proof will be based on a polynomial-time reduction of $\mathrm{EXT}(P)$ to a CS problem. Once this reduction is established, the **NP**-completeness of $\mathrm{EXT}(P)$ follows from results in [1] and [9].

Let M be a finite set of identities in the first order language of some algebra A. The set M is called *trivial* if there is a two element algebra similar to A whose basic operations are projections satisfying M. The set M is called *idempotent* if in any algebra the term operations satisfying M are idempotent. We say that an algebra B *admits a nontrivial Malcev condition* if there is a nontrivial set of identities satisfied by some term operations of B. We call an algebra *trivial* if its basic operations are projections. Let $n \geq 2$. An n-ary idempotent operation f is called a *Taylor operation* if it satisfies n identities of the form

$$f(\dots,\underbrace{x}_{i},\dots)=f(\dots,\underbrace{y}_{i},\dots),\quad i=1,2,\dots,n,$$

where x and y are the only variables occurring in the identities and $x \neq y$. The proof of the following theorem is due to Taylor; see Corollaries 5.2 and 5.3 in [17]. An alternate proof is given in [8]; see Theorem 9.4.

Theorem 4 (see [17]). *Let A be a idempotent algebra. Then the following are equivalent:*

(1) *A admits no nontrivial idempotent Malcev condition.*

(2) *There is a two element trivial algebra in the variety generated by A.*

(3) *There is no Taylor operation among the term operations of A.*

If A is a finite idempotent algebra satisfying the equivalent conditions of the preceding theorem, then a two element trivial algebra is a homomorphic image of a subalgebra of A^n for some finite n. In fact, more is true.

Proposition 5. *Let B be a two element algebra which is a homomorphic image of a subalgebra of a finite power of an idempotent algebra A. Then B is a homomorphic image of a subalgebra of A.*

Proof. Let C be a subalgebra of A^n, let h be a homomorphism of C onto B, and assume that $n>1$ is the least integer for which this situation holds. Let π_i denote the ith projection of A^n onto A, and let f_i denote the restriction of π_i to C. Let D denote the image of f_i. Suppose first that the kernel of f_i is contained in the kernel of h. Then by the second isomorphism theorem there exists a homomorphism of D onto B and this contradicts our choice of n. Hence there exists a pair (a,b) of elements of C such that $h(\{a,b\})=B$ and $a_i=b_i=c$ for some $c\in A$. However, since A is an idempotent algebra, the set X of all $x \in A^n$ such that $x_i = c$ is a subalgebra of A^n; in fact, it can clearly be embedded in A^{n-1} and it admits a homomorphism onto B, contradicting once more our choice of n. □

Let A be an algebra and R a relational set on the same base set. We say that *A is an algebra for R* or *R is a relational set for A* if the set of term operations of A coincides with the set of morphisms from finite powers of R to R. The following result is due to Jeavons; see Corollary 4.11 in [9].

THEOREM 6 (see [9]). *Let Q and R be finite relational sets on the same base set and let A be an algebra for Q. If R is of finite type and all the relations of R are subalgebras of finite powers of A, then CSP(R) reduces to CSP(Q) in polynomial time.*

The proof of the following result is basically contained in [1].

THEOREM 7. *Let R and Q be finite relational sets where R is of finite type. Suppose that A is an algebra on the same base set as R such that the relations of R are subalgebras of finite powers of A and that B is an algebra for Q. If A is a homomorphic image of a subalgebra of B, then there is a polynomial-time reduction of CSP(R) to CSP(Q).*

Proof. Let C be a subalgebra of B and φ a homomorphism from C onto A. Let R' be the relational set whose base set is the base set of Q and whose relations are the inverse images of the relations of R under φ^{-1} together with the unary relation C. It is a routine exercise to show that the relations of R' are preserved under the operations of B. So, it follows that the relations of R' are subalgebras of finite powers of B. Now, any instance X of CSP(R) is made into an instance Y of CSP(R') by adding the base set of X as the unary relation corresponding to C. Moreover, X is a satisfying instance of CSP(R) if and only if Y is a satisfying instance of CSP(R'). So CSP(R) reduces to CSP(R') in polynomial time. By the preceding theorem CSP(R') reduces to CSP(Q) in polynomial time, which concludes the proof. □

By a result of Schaefer [15], there exists a relational set R of finite type on the two element set such that CSP(R) is **NP**-complete. Combining this fact with the preceding theorems we get the following.

COROLLARY 8. *Let A be an idempotent algebra for a finite relational set Q. If A admits no nontrivial idempotent Malcev condition, then CSP(Q) is* **NP***-complete.*

Proof. Let R be a relational set of finite type on two elements such that CSP(R) is **NP**-complete. By Theorem 4 and Proposition 5, there exists a trivial two element algebra B which is a homomorphic image of a subalgebra of A. Since B is trivial, the relations of R are subalgebras of finite powers of B. Hence by Theorem 7 there is a polynomial-time reduction of CSP(R) to CSP(Q). □

We say that a poset P *admits a nontrivial idempotent Malcev condition* if $I(P)$ admits a nontrivial idempotent Malcev condition. Posets admitting a TSI operation are examples of such posets since any TSI operation is a Taylor operation. Next we show that if P admits no nontrivial Malcev condition, then EXT(P) is **NP**-complete. Let P_P be the relational set defined by equipping the base set of P with the order relation of P and all constants of P as one element unary relations. Clearly, P_P is a relational set for $I(P)$.

PROPOSITION 9. *The problems EXT(P) and CSP(P_P) are polynomial-time equivalent.*

Proof. First we reduce EXT(P) to CSP(P_P) in polynomial time. Let (Q,f) be an instance for EXT(P). By adding the unary relation $f^{-1}(p)$ for each $p \in P$ to the poset Q, we define an instance Q_P of CSP(P_P). Note that (Q,f) is extendible if and only if there is a morphism from Q_P to P_P.

Next we reduce $\mathrm{CSP}(P_P)$ to $\mathrm{EXT}(P)$ in polynomial time. Since P_P is of finite type we can assume without loss of generality that the instances of $\mathrm{CSP}(P_P)$ are relational sets similar to P_P. So let X be an arbitrary relational set similar to P_P. The reflexive, transitive closure of the binary relation of X is a quasi order, say α, on P. Let $\theta = \alpha \cap \alpha^{-1}$, the equivalence determined by α, as usual. Let $\overline{x}$ denote the θ-block of x for each $x \in X$. We associate a P-colored poset (H, f) with X, where $H = X/\theta$ and for each $x \in X$, $f(\overline{x}) = p$ if and only if x is in the unary relation of X that corresponds to p. So $f(\overline{x})$ is not defined whenever no unary relation of X contains x. In case that f is not well defined, X is not a satisfying instance of $\mathrm{CSP}(P_P)$. So we restrict ourselves to the case when f is well defined. Observe that the kernel of any morphism from X to P_P contains θ. So X is a satisfying instance of $\mathrm{CSP}(P_P)$ if and only if (H, f) is extendible. □

COROLLARY 10. *If P is a finite poset which admits no nontrivial Malcev condition, then EXT(P) is* **NP***-complete.*

Proof. By Proposition 9 it suffices to show that $\mathrm{CSP}(P_P)$ is **NP**-complete. Now, $I(P)$ is an idempotent algebra for P_P such that $I(P)$ admits no nontrivial idempotent Malcev condition. Then by Corollary 8 the problem $\mathrm{CSP}(P_P)$ is **NP**-complete. □

Posets admitting only projections as idempotent operations have been studied in numerous papers, e.g., [2], [4], [7], and [11]. Any poset P with an idempotent P-subalgebra that retracts onto one of these posets admits no nontrivial Malcev condition and hence yields an example where $\mathrm{EXT}(P)$ is **NP**-complete.

3. Concluding remarks. We saw that for the finite posets admitting a TSI operation of large enough arity the extendibility problem has polynomial-time complexity. On the other hand, we obtained that for finite posets which admit no nontrivial idempotent Malcev condition the extendibility problem is **NP**-complete.

Are there any finite posets not included in the two classes of posets mentioned in the preceding paragraph? There are CS problems without bounded width in **P** which are polynomial-time equivalent to the extendibility problem for a poset as proven in [6]. The anonymous referee, whom we thank for his/her judicious advice, noted that a resulting poset of that type might not have bounded width either. Nonetheless, the question still remains open.

Neither are we able to answer the following simpler question. Is there any finite poset P that admits a binary idempotent commutative operation but does not admit a TSI operation of arity $|P|$?

Acknowledgment. The authors would like to thank Pavol Hell for patiently mediating between the referee of the paper and the authors.

REFERENCES

[1] A. A. BULATOV, P. G. JEAVONS, AND A. A. KROKHIN, *Constraint satisfaction problems and finite algebras*, in Proceedings of the 27th International Colloquium on Automata, Languages and Programming, ICALP'00, Lecture Notes in Comput. Sci. 1853, Springer-Verlag, Berlin, 2000, pp. 272–282.

[2] E. COROMINAS, *Sur les ensembles ordonnés projectif et la properiétré du point fixe*, C. R. Acad. Sci. Paris Sér. I Math., 311 (1990), pp. 199–204.

[3] V. DALMAU AND J. PEARSON, *Closure functions and width* 1 *problems*, in Proceedings of the 5th International Conference on Principles and Practice of Constraint Programming, CP'99, Lecture Notes in Comput. Sci. 1713, Springer-Verlag, Berlin, 1999, pp. 159–173.

[4] B. A. DAVEY, J. B. NATION, R. N. MCKENZIE, AND P. P. PÁLFY, *Braids and their monotone clones*, Algebra Universalis, 32 (1994), pp. 153–176.

[5] D. Duffus and T. Goddard, *The complexity of the fixed point property*, Order, 13 (1996), pp. 209–218.

[6] T. Feder and M. Y. Vardi, *The computational structure of monotone monadic SNP and constraint satisfaction: A study through datalog and group theory*, SIAM J. Comput., 28 (1998), pp. 57–104.

[7] S. Hazan, *Two properties of projective orders*, Order, 9 (1992), pp. 233–238.

[8] D. Hobby and R. McKenzie, *The Structure of Finite Algebras*, Contemp. Math. 76, AMS, Providence, RI, 1988.

[9] P. Jeavons, *On the algebraic structure of combinatorial problems*, Theoret. Comput. Sci., 200 (1998), pp. 185–204.

[10] G. Kun and Cs. Szabó, *Order varieties and monotone retractions of finite posets*, Order, 18 (2001), pp. 79–88.

[11] B. Larose, *Minimal automorphic posets and the projection property*, Internat. J. Algebra Comput., 5 (1995), pp. 65–80.

[12] B. Larose and L. Zádori, *Algebraic properties and dismantlability of finite posets*, Discrete Math., 163 (1997), pp. 89–99.

[13] V. Pratt and J. Tiuryn, *Satisfiability of inequalities in a poset*, Fund. Inform., 28 (1996), pp. 165–182.

[14] I. Rival, *A fixed point theorem for finite partially ordered sets*, J. Combin. Theory Ser. A, 21 (1976), pp. 309–318.

[15] T. J. Schaefer, *The complexity of the satisfiability problems*, in Proceedings of the 10th ACM Symposium on the Theory of Computing (STOC), 1978, pp. 216–226.

[16] Cs. Szabó and L. Zádori, *Idempotent totally symmetric operations on finite posets*, Order, 18 (2001), pp. 39–47.

[17] W. Taylor, *Varieties obeying homotopy laws*, Canad. J. Math., 29 (1977), pp. 498–527.

SIAM J. Discrete Math.
Vol. 17, No. 1, pp. 122–133

COOPERATIVE GAMES UNDER AUGMENTING SYSTEMS[*]

JESÚS MARIO BILBAO[†]

Abstract. The goal of this paper is to develop a theoretical framework in order to analyze cooperative games in which only certain coalitions are allowed to form. We will axiomatize the structure of such allowable coalitions using the theory of *antimatroids*, a notion developed for combinatorially abstract sets. There have been previous models developed to confront the problem of unallowable coalitions. Games restricted by a communication graph were introduced by Myerson and Owen. We introduce a new combinatorial structure called *augmenting system*, which is a generalization of the antimatroid structure and the system of connected subgraphs of a graph. The main result of the paper is a direct formula of Shapley and Banzhaf values for games under augmenting systems restrictions.

Key words. cooperative game, Shapley value, Banzhaf value, set systems

AMS subject classification. 91A12

DOI. 10.1137/S0895480102402745

1. Introduction.

Cooperative games under combinatorial restrictions are cooperative games in which the players have restricted communication possibilities, which are defined by a combinatorial structure. The first model in which the restrictions are defined by the connected subgraphs of a graph is introduced by Myerson [11]. Since then, many other situations where players have communication restrictions have been studied in cooperative game theory. Contributions on graph-restricted games include Owen [12], Borm, Owen, and Tijs [3], and Hamiache [8]. In these models the possibilities of coalition formation are determined by the positions of the players in a *communication graph.* Another type of combinatorial structure introduced by Gilles, Owen, and van den Brink [7] is equivalent to a subclass of antimatroids. This line of research focuses on the possibilities of coalition formation determined by the positions of the players in the so-called *permission structure.* Sandholm et al. [14] analyze coalition formation in combinatorial problems.

In the present paper, we use the restricted cooperation model derived from a combinatorial structure called *augmenting system.* Section 2 introduces this structure, which is a generalization of the antimatroid structure and the system of connected subgraphs of a graph. Furthermore, this new set system includes the conjunctive and disjunctive systems derived from a permission structure. Section 3 introduces games under augmenting systems which generalize the ones studied on graphs and permission structures. Using the structural properties from these systems we will be able to express the dividends in terms of the original game. This result will be essential in section 4 to provide direct formulas to compute the Shapley and Banzhaf values for games under augmenting systems restrictions. In these formulas, these values are computed by means of the original game without having to calculate the restricted game and taking into account only the coalitions in the augmenting system. Finally, in section 5 we consider the potential and the Owen multilinear extension (MLE) for the restricted game. These results generalize, unify and simplify results of Owen [12],

[*]Received by the editors February 15, 2002; accepted for publication (in revised form) May 11, 2003; published electronically October 2, 2003.
http://www.siam.org/journals/sidma/17-1/40274.html

[†]Department of Applied Mathematics II, Escuela Superior de Ingenieros, Camino de los Descubrimientos, 41092 Sevilla, Spain (mbilbao@us.es, http://www.esi2.us.es/~mbilbao/).

Gilles, Owen, and van den Brink [7], and Bilbao [2].

2. Augmenting systems. Antimatroids were introduced by Dilworth [5] as particular examples of semimodular lattices. Since then, several authors have obtained the same concept by abstracting various combinatorial situations (see Korte, Lovász, and Schrader [10]). In this section, a general cooperation structure is introduced, which is a weakening of the antimatroid structure.

Let N be a finite set. A *set system* over N is a pair $(N,\mathcal{F})$ where $\mathcal{F}\subseteq 2^N$ is a family of subsets. The sets belonging to $\mathcal{F}$ are called *feasible.* We will write $S\cup i$ and $S\setminus i$ instead of $S\cup\{i\}$ and $S\setminus\{i\}$, respectively.

Definition 2.1. *A set system $(N,\mathcal{A})$ is an antimatroid if*

A1. $\emptyset\in\mathcal{A}$,

A2. *for $S,T\in\mathcal{A}$, we have $S\cup T\in\mathcal{A}$,*

A3. *for $S\in\mathcal{A}$ with $S\neq\emptyset$, there exists $i\in S$ such that $S\setminus i\in\mathcal{A}$.*

The definition of antimatroid implies the following *augmentation property*: If $S,T\in\mathcal{A}$ with $|T|>|S|$, then there exists $i\in T\setminus S$ such that $S\cup i\in\mathcal{A}$. We call a set system $(N,\mathcal{F})$ *normal* if $N=\bigcup_{S\in\mathcal{F}}S$. If $(N,\mathcal{A})$ is a normal antimatroid, then property A2 implies that $N\in\mathcal{A}$.

Definition 2.2. *An augmenting system is a normal set system $(N,\mathcal{F})$ with the following properties:*

P1. $\emptyset\in\mathcal{F}$,

P2. *for $S,T\in\mathcal{F}$ with $S\cap T\neq\emptyset$, we have $S\cup T\in\mathcal{F}$,*

P3. *for $S,T\in\mathcal{F}$ with $S\subset T$, there exists $i\in T\setminus S$ such that $S\cup i\in\mathcal{F}$.*

Remark. It follows from the definition that normal antimatroids are always augmenting systems.

Proposition 2.3. *An augmenting system $(N,\mathcal{F})$ is an antimatroid if and only if $\mathcal{F}$ is closed under union.*

Proof. The necessary condition follows from A2. Conversely, we only have to prove A3. Let $S\in\mathcal{F}$ with $S\neq\emptyset$. By property P3 there exists a chain of feasible subsets

$$\emptyset=S_0\subset S_1\subset\cdots\subset S_{s-1}\subset S_s=S$$

such that $S_k\in\mathcal{F}$ and $|S_k|=k$ for $0\leq k\leq s$. Hence there exists an element $i\in S$ such that $S\setminus i=S_{s-1}\in\mathcal{F}$. □

Example. The following collections of subsets of $N=\{1,\ldots,n\}$, given by $\mathcal{F}=2^N$ and $\mathcal{F}=\{\emptyset,\{1\},\ldots,\{n\}\}$, are the maximum augmenting system and a minimal augmenting system over N, respectively.

Example. In a communication graph $G=(N,E)$, the set system $(N,\mathcal{F})$ given by $\mathcal{F}=\{S\subseteq N:(S,E(S))$ is a connected subgraph of $G\}$ is an augmenting system.

Example. Gilles, Owen, and van den Brink [7] showed that the feasible coalitions system $(N,\mathcal{F})$ derived from the conjunctive or disjunctive approach contains the empty set and the ground set N and that it is closed under union. Algaba et al. [1] showed that the coalitions systems derived from the conjunctive and disjunctive approach were identified to *poset antimatroids* and *antimatroids with the path property*, respectively. Thus, these coalitions systems are augmenting systems.

Convex geometries are a combinatorial abstraction of convex sets introduced by Edelman and Jamison [6].

Definition 2.4. *A set system $(N,\mathcal{G})$ is a convex geometry if it satisfies the following properties:*

C1. $\emptyset\in\mathcal{G}$,

C2. *for $S, T \in \mathcal{G}$, we have $S \cap T \in \mathcal{G}$,*

C3. *for $S \in \mathcal{G}$ with $S \neq N$, there exists $i \in N \setminus S$ such that $S \cup i \in \mathcal{G}$.*

PROPOSITION 2.5. *An augmenting system $(N, \mathcal{F})$ is a convex geometry if and only if $\mathcal{F}$ is closed under intersection and $N \in \mathcal{F}$.*

Proof. The necessary conditions follow from properties C2 and C3. To prove sufficiency, note that $(N, \mathcal{F})$ satisfies C1 and C2, i.e., it is a closure system over N. Moreover, $(N, \mathcal{F})$ satisfies property P3 and $N \in \mathcal{F}$. Then for every $S \in \mathcal{F}$ with $S \neq N$, there exists $i \in N \setminus S$ such that $S \cup i \in \mathcal{F}$. □

DEFINITION 2.6. *Let $(N, \mathcal{F})$ be an augmenting system. For a feasible coalition $S \in \mathcal{F}$, we define the set $S^* = \{i \in N \setminus S : S \cup i \in \mathcal{F}\}$ of augmentations of S and the set $S^+ = S \cup S^* = \{i \in N : S \cup i \in \mathcal{F}\}$.*

PROPOSITION 2.7. *Let $(N, \mathcal{F})$ be an augmenting system. Then the interval $[S, S^+]_{\mathcal{F}} = \{C \in \mathcal{F} : S \subseteq C \subseteq S^+\}$ is a Boolean algebra for every nonempty $S \in \mathcal{F}$.*

Proof. It is suffices to show that $[S, S^+]_{\mathcal{F}} = \{C \subseteq N : S \subseteq C \subseteq S^+\}$, i.e., for every $C \subseteq N$ such that $S \subseteq C \subseteq S^+$ we have $C \in \mathcal{F}$. If $S^* = \emptyset$, then $[S, S^+]_{\mathcal{F}} = \{S\}$. Otherwise, $S^* = \{i_1, \ldots, i_p\}$ and $S \subseteq C \subseteq S^+$ implies $C = S \cup \{i_1, \ldots, i_q\}$ for some $1 \leq q \leq p$. We prove that $C \in \mathcal{F}$ by induction on q. For $q = 1$ we know that $S \cup \{i_1\} \in \mathcal{F}$. Assume $S \cup \{i_1, \ldots, i_k\} \in \mathcal{F}$. Since $S \cup \{i_{k+1}\} \in \mathcal{F}$ and $(S \cup \{i_1, \ldots, i_k\}) \cap (S \cup \{i_{k+1}\}) = S \neq \emptyset$, property P2 yields $S \cup \{i_1, \ldots, i_k, i_{k+1}\} \in \mathcal{F}$. □

Let $(N, \mathcal{F})$ be a set system and let $S \subseteq N$ be a subset. A feasible subset $C \in \mathcal{F}$ with $C \subseteq S$ is called a *basis* of S if $C \cup i \notin \mathcal{F}$ for all $i \in S \setminus C$. The maximal nonempty feasible subsets of S are called *components* of S. Clearly, every component of S is a basis of S. However, the converse is not true, as the following example shows.

Example. If $N = \{1, 2, 3\}$ and $\mathcal{F} = \{\emptyset, \{1\}, \{2\}, \{2, 3\}, N\}$, then $C = \{1\}$ is a basis of N, but the only component of N is the ground set N.

Observe that if $(N, \mathcal{A})$ is an antimatroid, then any subset $S \subseteq N$ has a unique basis given by the following operator $\text{int}(S) = \bigcup \{C \in \mathcal{A} : C \subseteq S\}$. This feasible set is also the unique component of S.

PROPOSITION 2.8. *Let $(N, \mathcal{F})$ be an augmenting system and let $S \subseteq N$ be a subset. Then a nonempty feasible subset $C \subseteq S$ is a basis of S if and only if C is a component of S.*

Proof. Let $C \in \mathcal{F}$ be a basis of S and suppose C is not a component of S, i.e., there exists $D \in \mathcal{F}$ such that $C \subset D \subseteq S$. Then because of P3 there exists $i \in D \setminus C \subseteq S \setminus C$ such that $C \cup i \in \mathcal{F}$, which is a contradiction. □

We denote by $C_{\mathcal{F}}(S)$ the set of the components of a subset $S \subseteq N$. Observe that the set $C_{\mathcal{F}}(S)$ may be the empty set. This set will play a role in the concept of a game restricted by an augmenting system.

PROPOSITION 2.9. *A set system $(N, \mathcal{F})$ satisfies property* P2 *if and only if for any $S \subseteq N$ with $C_{\mathcal{F}}(S) \neq \emptyset$, the components of S form a partition of a subset of S.*

Proof. We suppose that $(N, \mathcal{F})$ satisfies P2 and let S_1, S_2 be components of S. If $S_1 \cap S_2 \neq \emptyset$, then $S_1 \cup S_2 \in \mathcal{F}$ and we have that $S_i \subset S_1 \cup S_2 \subseteq S$ for $i \in \{1, 2\}$. This contradicts the fact that S_1 and S_2 are components of S. Conversely, assume for any S with $C_{\mathcal{F}}(S) \neq \emptyset$ that its components form a partition of a subset of S. Suppose that $(N, \mathcal{F})$ does not satisfy P2. Then there are $A, B \in \mathcal{F}$, with $A \cap B \neq \emptyset$ and $A \cup B \notin \mathcal{F}$. Hence there must be a component $C_1 \in C_{\mathcal{F}}(A \cup B)$ with $A \subseteq C_1$ and a component $C_2 \in C_{\mathcal{F}}(A \cup B)$ with $B \subseteq C_2$ such that $C_1 \neq C_2$. This contradicts the fact that the components of $A \cup B$ are disjoint. □

Let $N = \{1, \ldots, n\}$ be a set of players with $n > 2$ and we consider a subset S of starting players. If $i \in S$, then the set $\{i\}$ is feasible. Each starting player i looks for a player $k \notin S$ to generate a new feasible coalition $\{i, k\}$. These coalitions with cardinality 2 search for new players, which agree to join one by one. If we assume that common elements of two feasible coalitions are intermediaries between the two coalitions in order to establish the feasibility of its union, we obtain an augmenting system $(N, \mathcal{F})$. Since the individual players $k \notin S$ are not feasible, the family $\mathcal{F}$ is not generated by the connected subgraphs of a graph. Moreover, if players $i, j \in S$, then $\{i\}, \{j\} \in S$ and $\{i, j\} \notin S$ and hence $(N, \mathcal{F})$ is not an antimatroid.

Example. Let $N = \{1, 2, 3, 4\}$ and we consider $S_1 = \{1, 2, 4\}$ and $S_2 = \{1, 4\}$. By using the above coalition formation model we can obtain the following augmenting systems, represented in Figure 1.

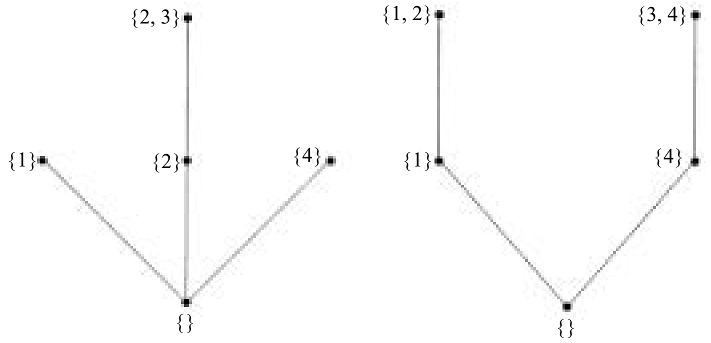


Fig. 1.

The sets of maximal feasible coalitions are partitions of the players into disjoint coalitions, that is, the coalition structures $CS_1 = \{\{1\}, \{4\}, \{2, 3\}\}$ and $CS_2 = \{\{1, 2\}, \{3, 4\}\}$. Coalition structure generation has been studied by Sandholm et al. [14].

Example. Let us consider $N = \{1, 2, 3, 4\}$ and

$$\mathcal{F} = \{\varnothing, \{1\}, \{4\}, \{1, 2\}, \{3, 4\}, \{1, 2, 3\}, \{2, 3, 4\}, N\}.$$

Since $\{1, 2, 3\}$ and $\{2, 3, 4\}$ are feasible, property P2 implies that the grand coalition N is a feasible set; see Figure 2.

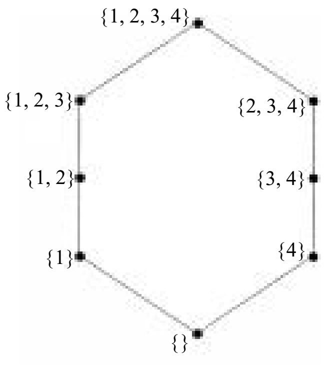


Fig. 2.

Example. The set system given by $N = \{1,2,3,4\}$ and

$$\mathcal{F} = \{\varnothing, \{1\}, \{4\}, \{1,2\}, \{1,3\}, \{2,4\}, \{3,4\}, \\ \{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\}, N\}$$

is an augmenting system. Since $\{1,4\} \notin \mathcal{F}$, the system $(N, \mathcal{F})$ represented in Figure 3 is not an antimatroid.

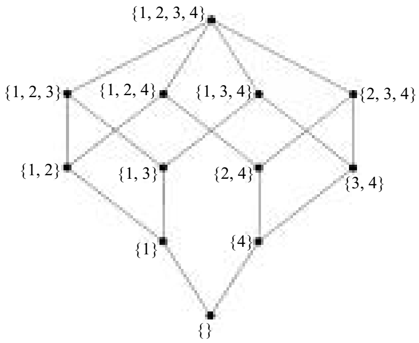


Fig. 3.

3. Games restricted by augmenting systems.

Definition 3.1. *Let $v : 2^N \to \mathbb{R}$ be a cooperative game and let $(N, \mathcal{F})$ be an augmenting system. The restricted game $v^{\mathcal{F}} : 2^N \to \mathbb{R}$ is defined by*

$$v^{\mathcal{F}}(S) = \sum_{T \in C_{\mathcal{F}}(S)} v(T).$$

Remark. If $(N, \mathcal{F})$ is the augmenting system given by the connected subgraphs of a graph $G = (N, E)$, then the game $\left(N, v^{\mathcal{F}}\right)$ is a graph-restricted game which is studied by Myerson [11] and Owen [12].

If $S \in \mathcal{F}$, then $v^{\mathcal{F}}(S) = v(S)$. Let us denote by Γ^N the vector space of all cooperative games (N, v), i.e., functions $v : 2^N \to \mathbb{R}$ such that $v(\emptyset) = 0$. Every cooperative game (N, v) is uniquely determined by the collection of its values $\{v(S) : S \subseteq N,\ S \neq \emptyset\}$. Then Γ^N will be identified with $\mathbb{R}^{2^n - 1}$. For any $S \subseteq N,\ S \neq \emptyset$, we define the *unanimity* game

$$u_S(T) = \begin{cases} 1 & \text{if } S \subseteq T, \\ 0 & \text{otherwise.} \end{cases}$$

Every game is a unique linear combination of unanimity games (cf. Shapley [15]),

$$v = \sum_{S \subseteq N} d_S u_S, \quad \text{where } d_S = \sum_{T \subseteq S} (-1)^{|S| - |T|} v(T).$$

We shall call d_S the *dividend* of S in the game v. Owen [12] showed the following property: *The unanimity games u_S, where S is connected in the graph G, form a basis of the graph-restricted games.*

Let $(N, \mathcal{F})$ be the system of connected subgraphs of a graph $G = (N, E)$. Hamiache [8] proved a formula for computing the dividends in the game $v^{\mathcal{F}}$ by using the values in the original game v. Next, we extend Hamiache's formula and Owen's property to the case when $(N,\mathcal{F})$ is an augmenting system.

PROPOSITION 3.2. *Let $(N,\mathcal{F})$ be an augmenting system and let (N,v) be a game. Then the restricted game $\left(N,v^{\mathcal{F}}\right)$ satisfies $v^{\mathcal{F}}=\sum_{C\in\mathcal{F}} d_C u_C$, where the dividend*

$$d_C=\sum_{\{S\in\mathcal{F}\,:\,S\subseteq C\subseteq S^+\}}(-1)^{|C|-|S|}v(S)$$

for every nonempty $C\in\mathcal{F}$ and $d_C=0$ otherwise.

Proof. The game $v^{\mathcal{F}}$ satisfies for every $C\subseteq N$

$$v^{\mathcal{F}}(C)=\sum_{T\subseteq N} d_T u_T(C)=\sum_{T\subseteq C} d_T,$$

where d_T the dividend of T in the game $v^{\mathcal{F}}$. Then, the Möbius inversion formula implies (see Stanley [16]) that

$$d_C=\sum_{T\subseteq C}(-1)^{|C|-|T|}v^{\mathcal{F}}(T).$$

It follows from $v^{\mathcal{F}}(\emptyset)=0$ that $d_\emptyset=0$. So we may assume that $C\neq\emptyset$. The definition of $v^{\mathcal{F}}$ implies that

$$\begin{aligned} d_C &= \sum_{T\subseteq C}(-1)^{|C|-|T|}\left(\sum_{S\in C_{\mathcal{F}}(T)} v(S)\right)\\ &= \sum_{\{S\in\mathcal{F}\,:\,S\subseteq C\}}\left(\sum_{\{T\subseteq C\,:\,S\in C_{\mathcal{F}}(T)\}}(-1)^{|C|-|T|}\right)v(S). \end{aligned}$$

Let $S\in\mathcal{F}$ with $S\subseteq C$. We first show that

$$\{T\subseteq C: S\in C_{\mathcal{F}}(T)\}=\left\{T\subseteq C: T\setminus S\subseteq C\setminus S^+\right\}.$$

We take $T\subseteq C$. If $S\in C_{\mathcal{F}}(T)$, then by Proposition 2.8, S is a basis of T and hence the set of its augmentations S^* satisfies $S^*\cap T=\emptyset$. Then for each $i\in T\setminus S$ we have $i\in C$ and $i\notin S\cup S^*=S^+$.

Conversely, let $T\subseteq C$ be a set such that $T\setminus S\subseteq C\setminus S^+$. Then for each $i\in T\setminus S$ we have $i\notin S^+$ and hence $S\cup i\notin\mathcal{F}$. Thus, the feasible set S is a basis of T and we conclude that $S\in C_{\mathcal{F}}(T)$.

Therefore, the coefficients of d_C satisfy

$$\begin{aligned} \sum_{\{T\subseteq C\,:\,S\in C_{\mathcal{F}}(T)\}}(-1)^{|C|-|T|} &= \sum_{\{T\subseteq C\,:\,S\subseteq T,\ T\setminus S\subseteq C\setminus S^+\}}(-1)^{|C|-|T|}\\ &= (-1)^{|C|-|S|}\left(\sum_{R\subseteq C\setminus S^+}(-1)^{-|R|}\right). \end{aligned}$$

Next, we compute

$$\sum_{R\subseteq C\setminus S^+}(-1)^{-|R|}=\sum_{R\subseteq C\setminus S^+}(-1)^{|R|}=(1-1)^{\left|C\setminus S^+\right|}=\begin{cases}1 & \text{if } C\setminus S^+=\emptyset,\\ 0 & \text{otherwise.}\end{cases}$$

Therefore, $C \setminus S^+ = \emptyset \Leftrightarrow C \subseteq S^+$, and hence

$$\begin{aligned} d_C &= \sum_{\{S \in \mathcal{F} \,:\, S \subseteq C,\ C \setminus S^+ = \emptyset\}} (-1)^{|C|-|S|} v\left(S\right) \\ &= \sum_{\{S \in \mathcal{F} \,:\, S \subseteq C \subseteq S^+\}} (-1)^{|C|-|S|} v(S). \end{aligned}$$

To complete the proof we observe that Proposition 2.7 implies that the set $C \in \mathcal{F}$. Otherwise $C \setminus S^+ \neq \emptyset$, and so $d_C = 0$ for all $C \notin \mathcal{F}$. □

4. The Shapley and Banzhaf values. Let (N, v) be a game and let $(N, \mathcal{F})$ be an augmenting system. The *Shapley value* for player i in the restricted game $v^{\mathcal{F}}$ is given by

$$\Phi_i\left(N, v^{\mathcal{F}}\right) = \sum_{\{S \subseteq N \,:\, i \in S\}} \frac{(s-1)!(n-s)!}{n!} \left[v^{\mathcal{F}}(S) - v^{\mathcal{F}}(S \setminus i)\right],$$

where $n = |N|$ and $s = |S|$. This value is an average of the *marginal contributions* $v^{\mathcal{F}}(S) - v^{\mathcal{F}}(S \setminus i)$ of a player i to all coalitions $S \in 2^N \setminus \{\emptyset\}$. In this value, the sets S of different size get different weight. The *Banzhaf value* for player i in the restricted game $v^{\mathcal{F}}$ is given by

$$\beta'_i\left(N, v^{\mathcal{F}}\right) = \sum_{\{S \subseteq N \,:\, i \in S\}} \frac{1}{2^{n-1}} \left[v^{\mathcal{F}}(S) - v^{\mathcal{F}}(S \setminus i)\right]$$

for all $i \in N$. If the number of players is n, then the function that measures the worst case running time for computing these indices is in $O\left(n2^n\right)$ (see Deng and Papadimitriou [4]). Moreover, to obtain the restricted game $v^{\mathcal{F}}$ we need to compute the set of the components $C_{\mathcal{F}}(S)$ of every subset $S \subseteq N$. Then it is necessary to consider all the feasible subsets of S, and hence the time complexity is $O\left(t\right)$, where

$$t = \sum_{s=0}^{n} \binom{n}{s} 2^s = 3^n.$$

The Shapley and Banzhaf values are linear mappings with respect to the characteristic function, and the images of the unanimity games are, respectively (cf. Owen [12]),

$$\Phi_i\left(N, u_S\right) = \begin{cases} 1\left/|S|\right. & \text{if } i \in S, \\ 0 & \text{otherwise,} \end{cases}$$

$$\beta'_i\left(N, u_S\right) = \begin{cases} 1\left/2^{|S \setminus i|}\right. & \text{if } i \in S, \\ 0 & \text{otherwise.} \end{cases}$$

In terms of dividends d_S in game $v^{\mathcal{F}}$, we have that

$$\begin{aligned} \Phi_i\left(N, v^{\mathcal{F}}\right) &= \sum_{\{S \subseteq N \,:\, i \in S\}} \frac{d_S}{|S|}, \\ \beta'_i\left(N, v^{\mathcal{F}}\right) &= \sum_{\{S \subseteq N \,:\, i \in S\}} \frac{d_S}{2^{|S \setminus i|}}. \end{aligned} \tag{1}$$

In the next theorem, two *explicit formulas,* in terms of v, for the Shapley and Banzhaf values of the players in the restricted game $v^{\mathcal{F}}$ are proved. These formulas generalize the results obtained by Bilbao [2] for games restricted by convex geometries.

THEOREM 4.1. *Let $(N,\mathcal{F})$ be an augmenting system and let (N,v) be a game. Then*

$$\Phi_i\left(N,v^{\mathcal{F}}\right)=\sum_{\{T\in\mathcal{F}\,:\,i\in T\}}\frac{(t-1)!\,t^*!}{t^+!}v(T)-\sum_{\{T\in\mathcal{F}\,:\,i\in T^*\}}\frac{t!\,(t^*-1)!}{t^+!}v(T),$$
$$\beta_i'\left(N,v^{\mathcal{F}}\right)=\sum_{\{T\in\mathcal{F}\,:\,i\in T\}}\frac{1}{2^{t^+-1}}v(T)-\sum_{\{T\in\mathcal{F}\,:\,i\in T^*\}}\frac{1}{2^{t^+-1}}v(T),$$

where $t=|T|$, $t^*=|T^*|$, and $t^+=|T^+|$.

Proof. By Proposition 3.2, we know that $d_S=0$ unless $S\in\mathcal{F}$. We use the formula (1) and Proposition 3.2 for computing

$$\begin{aligned}\Phi_i\left(N,v^{\mathcal{F}}\right)&=\sum_{\{S\in\mathcal{F}\,:\,i\in S\}}\frac{d_S}{|S|}\\&=\sum_{\{S\in\mathcal{F}\,:\,i\in S\}}\frac{1}{|S|}\left[\sum_{\{T\in\mathcal{F}\,:\,T\subseteq S\subseteq T^+\}}(-1)^{|S|-|T|}v(T)\right].\end{aligned}$$

Reversing the order of summation and denoting $s=|S|$ and $t=|T|$, we obtain

$$\begin{aligned}\Phi_i\left(N,v^{\mathcal{F}}\right)&=\sum_{T\in\mathcal{F}}\left[\sum_{\{S\in\mathcal{F}\,:\,i\in S,\ T\subseteq S\subseteq T^+\}}\frac{(-1)^{s-t}}{s}\right]v(T)\\&=\sum_{T\in\mathcal{F}}c_i(T)v(T),\end{aligned}$$

where

$$c_i(T)=\sum_{\{S\in\mathcal{F}\,:\,T\cup i\subseteq S\subseteq T^+\}}\frac{(-1)^{s-t}}{s}.$$

First, we suppose $i\in T$. By Proposition 2.7 the interval $[T,T^+]$ is a Boolean algebra and hence the summation index is $\{S\subseteq N:T\subseteq S\subseteq T^+\}$. Now we consider $S=T\cup R$, where $R=S\setminus T$, $r=|R|$, and $t^*=|T^*|$. Then

$$\begin{aligned}c_i(T)&=\sum_{R\subseteq T^*}\frac{(-1)^r}{t+r}=\sum_{r=0}^{t^*}\binom{t^*}{r}\frac{(-1)^r}{t+r}\\&=\sum_{r=0}^{t^*}\binom{t^*}{r}(-1)^r\int_0^1 x^{t+r-1}\,dx\\&=\int_0^1 x^{t-1}\sum_{r=0}^{t^*}\binom{t^*}{r}(-x)^r\,dx\\&=\int_0^1 x^{t-1}(1-x)^{t^*}\,dx\\&=\frac{(t-1)!\,t^*!}{t^+!}.\end{aligned}$$

Next, assume that $i \notin T$; hence the index is $\{S \in \mathcal{F} : T \cup i \subseteq S \subseteq T^+\}$. Then $i \in T^+ \setminus T$ and hence $i \in T^*$. Now the previous result yields (note that $[T \cup i, T^+]$ is a Boolean algebra)

$$c_i(T) = -\sum_{\{S \subseteq N \,:\, T \cup i \subseteq S \subseteq T^+\}} \frac{(-1)^{s-(t+1)}}{s} = -\frac{t!(t^* - 1)!}{t^+!}.$$

Inserting the coefficients, we have

$$\Phi_i \left(N, v^{\mathcal{F}}\right) = \sum_{\{T \in \mathcal{F} \,:\, i \in T\}} \frac{(t-1)!\, t^*!}{t^+!} v(T) - \sum_{\{T \in \mathcal{F} \,:\, i \in T^*\}} \frac{(t)!(t^* - 1)!}{t^+!} v(T). \tag{2}$$

The proof of the formula of the Banzhaf value is similar. The only difference is that the coefficients are

$$c_i(T) = \sum_{r=0}^{t^*} \binom{t^*}{r} (-1)^r \left(\frac{1}{2}\right)^{t+r-1} = \left(\frac{1}{2}\right)^{t^+-1} \quad \text{if } i \in T,$$

$$c_i(T) = -\left(\frac{1}{2}\right)^{t^+-1} \quad \text{if } i \in T^*. \qquad \square$$

Remark. Notice that if $\mathcal{F} = 2^N$, then $T^* = N \setminus T$ and $T^+ = N$ for every $T \in \mathcal{F}$. Thus, the formulas obtained in the above theorem are equal to the classical Shapley and Banzhaf values for the game v. Moreover, equation (2) is equal to the equation of Shapley [15].

Let us consider a set system $(N, \mathcal{F})$. An element i of a feasible set $S \in \mathcal{F}$ is an *extreme point* of S if $S \setminus i \in \mathcal{F}$. The set of extreme points of S is denoted by $\mathrm{ex}(S)$. The formulas for computing the Shapley and Banzhaf values of the players in the restricted game $v^{\mathcal{F}}$ can be further simplified when the player is an extreme point of every feasible coalition. Before doing so, we will need a lemma.

Lemma 4.2. *Let $(N, \mathcal{F})$ be an augmenting system. If $i \in \mathrm{ex}(S)$ for all $S \in \mathcal{F}$ which contains i with $S \neq \{i\}$, then $(S \setminus i)^+ = S^+$.*

Proof. Note first that $i \in (S \setminus i)^+$ and $i \in S^+$. For every $j \in (S \setminus i)^+$ with $j \neq i$, we have $(S \setminus i) \cup j \in \mathcal{F}$. Then $((S \setminus i) \cup j) \cap S = S \setminus i \neq \emptyset$ implies $((S \setminus i) \cup j) \cup S = S \cup j \in \mathcal{F}$ and hence $j \in S^+$. Conversely, for every $j \in S^+$, $j \neq i$, we know that $S \cup j \in \mathcal{F}$. Since $i \in S \subseteq S \cup j$, the assumption implies that $i \in \mathrm{ex}(S \cup j)$. Then $(S \cup j) \setminus i = (S \setminus i) \cup j \in \mathcal{F}$ and thus $j \in (S \setminus i)^+$. $\square$

Theorem 4.3. *Let $(N, \mathcal{F})$ be an augmenting system and let (N, v) be a game such that $v(i) = 0$ for all $i \in N$. If $i \in \mathrm{ex}\,(S)$ for all $S \in \mathcal{F}$ that contains i, then*

$$\Phi_i \left(N, v^{\mathcal{F}}\right) = \sum_{\{S \in \mathcal{F} \,:\, i \in S,\ |S| > 1\}} \frac{(s-1)!\, s^*!}{s^+!} \left[v\left(S\right) - v(S \setminus i)\right],$$

$$\beta_i' \left(N, v^{\mathcal{F}}\right) = \sum_{\{S \in \mathcal{F} \,:\, i \in S,\ |S| > 1\}} \frac{1}{2^{s^+-1}} \left[v(S) - v(S \setminus i)\right],$$

where $s = |S|$, $s^* = |S^*|$, and $s^+ = |S^+|$.

Proof. We remark first that if i satisfies the hypothesis, then

$$\{S \in \mathcal{F} : i \in S,\ |S| > 1\} = \{S \in \mathcal{F} : i \in \mathrm{ex}\,(S)\,,\ |S| > 1\}\,.$$

Taking $T = S \setminus i$ we obtain $\{T \in \mathcal{F} : i \in T^*\} = \{S \setminus i : S \in \mathcal{F},\ i \in \mathrm{ex}\,(S)\}$. Next, we apply Theorem 4.1 and therefore, by Lemma 4.2,

$$\begin{aligned}
\Phi_i\left(N, v^{\mathcal{F}}\right) &= \sum_{\{S \in \mathcal{F} \,:\, i \in S\}} \frac{(s-1)!\, s^*!}{s^+!} v(S) - \sum_{\{T \in \mathcal{F} \,:\, i \in T^*\}} \frac{t!\,(t^*-1)!}{t^+!} v(T) \\
&= \sum_{\{S \in \mathcal{F} \,:\, i \in \mathrm{ex}(S),\ |S|>1\}} \frac{(s-1)!\, s^*!}{s^+!} v(S) \\
&\quad - \sum_{\{S \in \mathcal{F} \,:\, i \in \mathrm{ex}(S),\ |S|>1\}} \frac{(s-1)!\, s^*!}{s^+!} v(S \setminus i) \\
&= \sum_{\{S \in \mathcal{F} \,:\, i \in S,\ |S|>1\}} \frac{(s-1)!\, s^*!}{s^+!} \left[v\left(S\right) - v(S \setminus i)\right]
\end{aligned}$$

(note that $v(i) = 0$ for all $i \in N$). The result for the Banzhaf value follows similarly. □

Remark. Let $(N, \mathcal{F})$ be an augmenting system that is a convex geometry. Then for every $i \in \mathrm{ex}\,(N)$ we have $S \setminus i = (N \setminus i) \cap S \in \mathcal{F}$ for all $S \in \mathcal{F}$ such that $i \in S$. Hence, if $i \in \mathrm{ex}\,(N)$, then $i \in \mathrm{ex}\,(S)$ for all $S \in \mathcal{F}$ with $i \in S$.

Example. Let $K_{1,n-1}$ be a star on n vertices and let 1 be the center of star. The augmenting system of the connected subgraphs of $K_{1,n-1}$ is given by $\mathcal{F} = \{S \subseteq N : 1 \in S \text{ or } |S| = 1\}$. Then $\mathrm{ex}(N) = \{2, \ldots, n\}$, and for all $S \in \mathcal{F}$ such that $|S| > 1$, we infer that $1 \in S$, $S^* = N \setminus S$, and $S^+ = N$. Moreover, the set $\{S \in \mathcal{F} : 1 \in S^*,\ |S| > 1\} = \emptyset$. Using these properties, the following results can be derived from Theorems 4.1 and 4.3:

1. If (N, v) is a game such that $v(i) = 0$ for all $i \in N$, then

$$\Phi_1\left(N, v^{\mathcal{F}}\right) = \sum_{\{S \in \mathcal{F} \,:\, 1 \in S,\ |S|>1\}} \frac{(s-1)!(n-s)!}{n!} v(S).$$

2. If (N, v) is a game such that $v(i) = 0$ for all $i \in N$, then

$$\Phi_i\left(N, v^{\mathcal{F}}\right) = \sum_{\{S \in \mathcal{F} \,:\, i \in S,\ |S|>1\}} \frac{(s-1)!(n-s)!}{n!} \left[v(S) - v(S \setminus i)\right]$$

for all $i \in \{2, \ldots, n\}$.

Remark. The time complexity of the direct formulas showed in Theorems 4.1 and 4.3 is polynomial in the cardinality $|\mathcal{F}|$.

Example. Let us consider an augmenting system $(N, \mathcal{F})$ such that the family of its maximal elements is a coalition structure $CS = \{T_1, \ldots, T_p\}$. Then the number of feasible elements is

$$|\mathcal{F}| = |T_1| + \cdots + |T_p| + 1 = |N| + 1,$$

and hence $|\mathcal{F}|$ is polynomial in $|N|$. For instance, the augmenting systems represented in Figure 1 satisfy $|\mathcal{F}| = 5$.

Example. Let $(N, \mathcal{F})$ be an augmenting system with exactly two maximal chains. Then $|\mathcal{F}| = 2\,(|N| - 1) + 2 = 2\,|N|$, and hence $|\mathcal{F}|$ is polynomial in $|N|$. For instance, the augmenting system represented in Figure 2 satisfies $|\mathcal{F}| = 8$.

5. The potential and the MLE. The potential function for cooperative games was defined by Hart and Mas-Colell [9]. Given a game (N, v) and a coalition $S \subseteq N$, the *subgame* (S, v) is obtained by restricting v to 2^S. Let Γ denote the set of all games. The *potential* is a function $P : \Gamma \to \mathbb{R}$ which assigns to each game (N, v) a real number $P(N, v)$ and satisfies the following recursive equations:

$$P(\emptyset, v) = 0, \quad P(S, v) = \frac{1}{|S|}\left[v(S) + \sum_{i \in S} P\left(S \setminus i, v\right)\right]$$

for all nonempty $S \subseteq N$. Then the *marginal contribution* of i coincides with its Shapley value $P\left(N, v\right) - P\left(N \setminus i, v\right) = \Phi_i\left(N, v\right)$ for all $i \in N$. Moreover, there are two explicit formulas for the potential:

$$P(N, v) = \sum_{S \subseteq N} \frac{d_S}{|S|}, \quad P(N, v) = \sum_{S \subseteq N} \frac{(s-1)!(n-s)!}{n!} v(S),$$

where $s = |S|$ and $n = |N|$. The explicit formula for the potential of $v^{\mathcal{F}}$ can be obtained by a method similar to the one that is used in Theorem 4.1.

THEOREM 5.1. *Let $(N, \mathcal{F})$ be an augmenting system and let (N, v) be a game. Then*

$$P\left(N, v^{\mathcal{F}}\right) = \sum_{S \in \mathcal{F}} \frac{(s-1)!\, s^*!}{s^+!} v(S),$$

where $s = |S|$, $s^* = |S^*|$ and $s^+ = |S^+|$.

The *MLE* of the game (N, v) is the function of n real variables (see Owen [13]) $f(v)\left(q_1, \ldots, q_n\right) = \sum_{S \subseteq N} \prod_{j \in S} q_j \; d_S$, where d_S is the dividend of S in the game (N, v). Owen showed that

$$\Phi_i(N, v) = \int_0^1 \frac{\partial f(v)}{\partial q_i}\left(t, \ldots, t\right)\, dt,$$
$$\beta_i'(N, v) = \frac{\partial f(v)}{\partial q_i}\left(\frac{1}{2}, \ldots, \frac{1}{2}\right).$$

PROPOSITION 5.2. *Let $(N, \mathcal{F})$ be an augmenting system and let (N, v) be a game. Then the MLE of $v^{\mathcal{F}}$ is given by*

$$f\left(v^{\mathcal{F}}\right)\left(q_1, \ldots, q_n\right) = \sum_{S \in \mathcal{F}} \prod_{j \in S} q_j \left(\sum_{\{T \in \mathcal{F} \,:\, T \subseteq S \subseteq T^+\}} (-1)^{|S|-|T|} v(T)\right).$$

REFERENCES

[1] E. ALGABA, J. M. BILBAO, R. VAN DEN BRINK, AND A. JIMÉNEZ-LOSADA, *Cooperative games on antimatroids,* Center discussion paper 124, Tilburg University, The Netherlands, 2000; Discrete Math., submitted.

[2] J. M. BILBAO, *Values and potential of games with cooperation structure,* Internat. J. Game Theory, 27 (1998), pp. 131–145.

[3] P. BORM, G. OWEN, AND S. TIJS, *On the position value for communication situations,* SIAM J. Discrete Math., 5 (1992), pp. 305–320.

[4] X. DENG AND C. H. PAPADIMITRIOU, *On the complexity of cooperative solution concepts*, Math. Oper. Res., 19 (1994), pp. 257–266.

[5] R. P. Dilworth, *Lattices with unique irreducible decompositions,* Ann. Math., 41 (1940), pp. 771–777.

[6] P. H. Edelman and R. E. Jamison, *The theory of convex geometries,* Geom. Dedicata, 19 (1985), pp. 247–270.

[7] R. P. Gilles, G. Owen, and R. van den Brink, *Games with permission structures: The conjunctive approach,* Internat. J. Game Theory, 20 (1992), pp. 277–293.

[8] G. Hamiache, *A value with incomplete communication,* Games Economic Behav., 26 (1999), pp. 59–78.

[9] S. Hart and A. Mas-Colell, *The potential of the Shapley value,* in The Shapley Value: Essays in Honor of Lloyd S. Shapley, Cambridge University Press, Cambridge, UK 1988, pp. 127–137.

[10] B. Korte, L. Lovász, and R. Schrader, *Greedoids,* Springer, Berlin, 1991.

[11] R. B. Myerson, *Graphs and cooperation in games,* Math. Oper. Res., 2 (1977), pp. 225–229.

[12] G. Owen, *Values of graph-restricted games,* SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 210–220.

[13] G. Owen, *Multilinear extension of games,* in The Shapley Value: Essays in Honor of Lloyd S. Shapley, Cambridge University Press, Cambridge, UK 1988, pp. 139–151.

[14] T. Sandholm, K. Larson, M. Andersson, O. Shehory, and F. Tohmé, *Coalition structure generation with worst case guarantees*, Artificial Intelligence, 111 (1999), pp. 209–238.

[15] L. S. Shapley, *A value for n-person games,* in Contributions to the Theory of Games II, Ann. of Math. Stud. 28, Princeton University Press, Princeton, NJ, 1953, pp. 307–317.

[16] R. P. Stanley, *Enumerative Combinatorics* I*,* Wadsworth, Monterey, CA, 1986.

SIAM J. Discrete Math.
Vol. 17, No. 1, pp. 134–160

COMPARING TOP k LISTS[*]

RONALD FAGIN[†], RAVI KUMAR[†], AND D. SIVAKUMAR[†]

Abstract. Motivated by several applications, we introduce various distance measures between "top k lists." Some of these distance measures are metrics, while others are not. For each of these latter distance measures, we show that they are "almost" a metric in the following two seemingly unrelated aspects:

(i) they satisfy a relaxed version of the polygonal (hence, triangle) inequality, and

(ii) there is a metric with positive constant multiples that bound our measure above and below.

This is not a coincidence—we show that these two notions of almost being a metric are the same. Based on the second notion, we define two distance measures to be *equivalent* if they are bounded above and below by constant multiples of each other. We thereby identify a large and robust equivalence class of distance measures.

Besides the applications to the task of identifying good notions of (dis)similarity between two top k lists, our results imply polynomial-time constant-factor approximation algorithms for the *rank aggregation problem* with respect to a large class of distance measures.

Key words. triangle inequality, polygonal inequality, metric, near metric, distance measures, top k list, rank aggregation

AMS subject classifications. 68R05, 68W25, 54E99

DOI. S0895480102412856

1. Introduction.

The notion of a "top k list" is ubiquitous in the field of information retrieval (IR). A top 10 list, for example, is typically associated with the "first page" of results from a search engine. While there are several standard ways for *measuring* the "top k quality" of an IR system (e.g., precision and recall at various values of k), it appears that there is no well-studied and well-understood method for *comparing* two top k lists for similarity/dissimilarity. Methods based on precision and recall yield a way to compare two top k lists by comparing them both to "ground truth." However, there are two limitations of such approaches: First, these methods typically give absolute (unary) ratings of top k lists, rather than give a relative, binary measure of distance. Second, for IR in the context of the world-wide web, there is often no clear notion of what ground truth is, so precision and recall are harder to use.

These observations lead to the following question in discrete mathematics: *How do we define reasonable and meaningful distance measures between top k lists*? We motivate the study of this problem by sketching some applications.

Applications. The first group of applications we describe is in the comparison of various search engines, or of different variations of the same search engine. What could be a more natural way to compare two search engines than by comparing their visible outputs (namely, their top k lists)? It is also important to compare variations (using slightly different ranking functions) of the same search engine as an aid in the design of ranking functions. In particular, we can use our methodology to test the effect on the

[*] Received by the editors July 23, 2002; accepted for publication (in revised form) September 8, 2002; published electronically October 14, 2003. A preliminary version of this paper appeared in *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM, New York, SIAM, Philadelphia, 2003, pp. 28–36.

http://www.siam.org/journals/sidma/17-1/41285.html

[†] IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120 (fagin@almaden.ibm.com, ravi@almaden.ibm.com, siva@almaden.ibm.com).

top k lists of adding/deleting ranking heuristics to/from the search engine. Similar issues include understanding the effect of augmenting the "crawl" data to add more documents, of indexing more data types (e.g., PDF documents), etc. For a more complex application in this group, consider a large-scale search engine. Typically, its ranking function is a composite algorithm that builds on several simpler ranking functions, and the following questions are of interest: What is the "contribution" of each component to the final ranking algorithm (i.e., how similar is the top k composite output to the top k of each of its components), and how similar is each component to the others? A good quantitative way to measure these (which our methodology supplies) could be a valuable tool in deciding which components to retain, enhance, or delete so as to design a better ranking algorithm. Similarly, our methodology can be used to compare a "metasearch" engine with each of its component search engines in order to understand the degree to which the metasearch engine aligns itself with each of its components. In section 9, we report our results on the comparisons of seven popular Web search engines and on comparing a metasearch engine with its components.

The second group of the applications can be classified as "engineering optimizations." A fairly simple example is a system that draws its search results from several servers; for the sake of speed, a popular heuristic is to send the query to the servers and return the responses as soon as, say, 75% of the servers have responded. Naturally, it is important to ensure that the quality of the results are not adversely affected by this approximation. What one needs here are meaningful and quantitative measures with which to estimate the difference in the top k lists caused by the approximation. A more subtle example in the same category is the following (where our methodology has already been successfully utilized). Carmel et al. [CCF$^+$01] explored the effect of pruning the index information of a search engine. Their experimental hypothesis, which they verified using one of our distance measures, was that their pruning technique would have only small effects on the top k list for moderate values of k.[1] Since what a user sees is essentially a top k list, they concluded that they could prune the index greatly, which resulted in better space and time performance, without much effect on the search results. Kamvar et al. [KHMG03] have used one of our distance measures in evaluating the quality of an approximate version of the PageRank ranking algorithm. Another scenario in a similar vein is in the area of approximate near-neighbor searching, a very common technique for categorization problems. Here an important goal is to understand the difference between approximate and exact near-neighbor search; once again, since what matters the most are the top few results, our problem arises naturally.

Another application of comparing top k lists arises from the processing of data logs to discover emerging trends (see [CCF02] for an example). For example, a search engine could compute the top 100 queries each day and see how they differ from day to day, from month to month, etc. Other examples include processing inventory logs and sales logs in retail stores, logs of stocks traded each day, etc. In these cases, a spike in the difference between day-to-day or hour-to-hour top k lists could trigger a closer analysis and action (e.g., buy/sell shares, add inventory, etc.). For these settings, one needs good notions of the difference between two given top k lists.

Finally, we consider the context of synthesizing a good composite ranking function from several simpler ones. In the *rank aggregation problem* [DKNS01], given

[1] In fact, our first author is a coauthor of [CCF$^+$01] and the need for comparing top k lists that arose in that paper is what led us to the research in this paper.

several top k lists, the goal is to find a top k list that is a "good" consolidation of the given lists. In [DKNS01] this problem is formulated by asking for an aggregation that has the minimum total distance with respect to the given lists, where the distance is computed according to some distance measure of interest. The choice of distance measure turns out to have a direct bearing on the complexity of computing the best solution: some distance measures lead to NP-hard optimization problems, while others admit polynomial-time solutions. A main algorithmic consequence of our work is in enabling the design of efficient constant-factor approximation algorithms for the aggregation problem with respect to a large class of distance measures. This is achieved by identifying a class of distance measures that are within constant factors of each other.

Results. We approach the problem of defining distance measures between top k lists from many angles. We make several proposals for distance measures, based on various motivating criteria—ranging from naive, intuitive ones to ones based on rigorous mathematics. While the plethora of measures is good news (since it gives a wide choice), it also poses the challenging question of how to understand their relative merits, or how to make a sound choice among the many competing proposals.

One of our main contributions is a unified framework in which to catalog and organize various distance measures. Concretely, we propose the notion of an *equivalence class* of distance measures and, in particular, we place many of the proposed distance measures into one large equivalence class (which we dub the "big equivalence class"). Our big equivalence class encompasses many measures that are intuitively appealing (but whose mathematical properties are nebulous), as well as ones that were derived via rigorous mathematics (but lacking in any natural, intuitive justification that a user can appreciate). The main message of the equivalence class concept is that up to constant factors (that do not depend on k), all distance measures in an equivalence class are essentially the same.

Our equivalence classes have the property that if even one distance measure in a class is a *metric* (in the usual mathematical sense), then each of the others in that class is a "near metric." To make the foregoing idea precise, we present two distinct but seemingly unrelated definitions of a near metric. The first says that it satisfies a relaxed version of the "polygonal inequality" (the natural extension of the standard triangle inequality). The second says that there exists a metric with positive constant multiples that bound our measure above and below. We prove the surprising result that these two notions of near metric are, in fact, equivalent.

Our results have the following two consequences:

(1) The task of choosing a distance measure for IR applications is now considerably simplified. The only conscious choice a user needs to make is about which equivalence class to use, rather than which distance measure to use. Our personal favorite is the big equivalence class that we have identified, mainly because of the rich variety of underlying intuition and the mathematically clean and algorithmically simple methods that it includes.

(2) We obtain constant-factor approximation algorithms for the rank aggregation problem with respect to every distance measure in our big equivalence class. This is achieved using the fact that the rank aggregation problem can be optimally solved in polynomial time (via minimum cost perfect matching) for one of the distance measures in this equivalence class.

As we noted, in section 9 we present an illustration of the applicability of our methods in the context of search and metasearch. Based on the results for 750 user queries, we study the similarities between the top 50 lists of seven popular Web search engines and also their similarity to the top 50 list of a metasearch engine built using the seven search engines. The quantitative comparison of the search engines' top 50 results brings some surprising qualitative facts to light. For example, our experiments reveal that AOL Search and MSN Search yield very similar results, despite the fact that these are competitors. Further analysis reveals that the crawl data for these search engines (and also for the search engine HotBot) comes in part from Inktomi. The fact that the top 50 results from HotBot are only moderately similar to that of AOL Search and MSN Search suggests that while they all use crawl data from Inktomi, HotBot probably uses a ranking function quite different from those of AOL and MSN. We believe these studies make an excellent case for the applicability of quantitative methods in comparing top k lists.

Methodology. A special case of a top k list is a "full list," that is, a permutation of all of the objects in a fixed universe. There are several standard methods for comparing two permutations, such as Kendall's tau and Spearman's footrule (see the textbooks [Dia88, KG90]). We cannot simply apply these known methods, since they deal only with comparing one permutation against another over the same elements. Our first (and most important) class of distance measures between top k lists is obtained by various natural modifications of these standard notions of distances between permutations.

A fairly straightforward attempt at defining a distance measure is to compute the intersection of the two top k lists (viewing them as sets). This approach has in fact been used in several papers in IR [Lee95, Lee97, CCF$^+$01]. In order to obtain a metric, we consider the notion of the symmetric difference (union minus the intersection), appropriately scaled. This, unfortunately, is not adequate for the top k distance problem, since two top 10 lists that are reverses of each other would be declared to be "very close." We propose natural extensions of this idea that leads to a metric for top k lists. Briefly, the idea is to truncate the top k lists at various points $i \leq k$, compute the symmetric difference metric between the resulting top i lists, and take a suitable combination of them. This gives a second type of notion of the distance between top k lists.

As we noted, our distance measure based on the intersection gives a metric. What about our distance measures that are generalizations of metrics on permutations? Some of these turn out to be metrics, but others do not. For each of these distance measures d that is not a metric, we show that d is a "near metric" in two seemingly different senses. Namely, d satisfies each of the following two properties.

Metric boundedness property. There is a metric d' and positive constants c_1 and c_2 such that for all x, y in the domain, $c_1 d'(x,y) \leq d(x,y) \leq c_2 d'(x,y)$ for all x, y in the domain.

Thus, metric boundedness says that d and some metric d' are within constant multiples of each other.

Relaxed polygonal inequality. There is a constant c such that for all $n > 1$ and $x, z, x_1, \ldots, x_{n-1}$ in the domain, $d(x,z) \leq c(d(x,x_1) + d(x_1,x_2) + \cdots + d(x_{n-1},z))$.

As remarked earlier, we show the surprising fact that these two seemingly unrelated notions of being a "near metric" are the same. Note that the relaxed polygonal inequality immediately implies the relaxed triangle inequality [FS98], which says that there is a constant c such that $d(x,z) \leq c(d(x,y)+d(y,z))$ for all x, y, z in the domain. Relaxed triangle and polygonal inequalities suggest that the notion of "closeness" under these measures are "reasonably transitive." Interestingly enough, the equivalence of our two notions of "near metric" requires that we consider the relaxed polygonal inequality rather than simply the relaxed triangle inequality; the relaxed triangle inequality is not sufficient to imply the metric boundedness property.

Organization. In section 2, we review two metrics on permutations, which form the basis for various distance measures that we define and study. In section 3, we develop our new distance measures between top k lists. In section 4, we present various notions of near metric, and show the equivalence between metric boundedness and the relaxed polygonal inequality. In section 5, we define the notion of equivalence of distance measures and show that all of our distance measures are in one large and robust equivalence class, called the "big equivalence class." Thus each of the distance measures between top k lists introduced in section 3 is a metric or a near metric. In section 6, we give an algorithmic application that exploits distance measures being in the same equivalence class. In section 7, we discuss two approaches based on Spearman's rho and symmetric difference. In section 8, we discuss the interpolation criterion—a natural and desirable property of a distance measure. In section 10, we conclude the paper.

2. Metrics on permutations.

The study of metrics on permutations is classical. The book by Kendall and Gibbons [KG90] provides a detailed account of various methods. Diaconis [Dia88] gives a formal treatment of metrics on permutations. We now review two well-known notions of metrics on permutations.

A *permutation* σ is a bijection from a set $D = D_\sigma$ (which we call the *domain*, or *universe*) onto the set $[n] = \{1, \ldots, n\}$, where n is the size $|D|$ of D. Let S_D denote the set of all permutations of D. For a permutation σ, we interpret $\sigma(i)$ as the position (or rank) of element i. We say that *i is ahead of j in σ* if $\sigma(i) < \sigma(j)$. Let $\mathcal{P} = \mathcal{P}_D = \{\{i,j\} \mid i \neq j \text{ and } i,j \in D\}$ be the set of unordered pairs of distinct elements. Let σ_1, σ_2 be two members of S_D.

Kendall's tau metric between permutations is defined as follows. For each pair $\{i,j\} \in \mathcal{P}$ of distinct members of D, if i and j are in the same order in σ_1 and σ_2, then let $\bar{K}_{i,j}(\sigma_1,\sigma_2) = 0$; if i and j are in the opposite order (such as i being ahead of j in σ_1 and j being ahead of i in σ_2), then let $\bar{K}_{i,j}(\sigma_1,\sigma_2) = 1$. Kendall's tau is given by $K(\sigma_1,\sigma_2) = \sum_{\{i,j\}\in\mathcal{P}} \bar{K}_{i,j}(\sigma_1,\sigma_2)$. The maximum value of $K(\sigma_1,\sigma_2)$ is $n(n-1)/2$, which occurs when σ_1 is the reverse of σ_2 (that is, when $\sigma_1(i) + \sigma_2(i) = n+1$ for each i). Kendall's tau turns out to be equal to the number of exchanges needed in a bubble sort to convert one permutation to the other.

Spearman's footrule metric is the L_1 distance between two permutations. Formally, it is defined by $F(\sigma_1,\sigma_2) = \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)|$. The maximum value of $F(\sigma_1,\sigma_2)$ is $n^2/2$ when n is even, and $(n+1)(n-1)/2$ when n is odd. As with Kendall's tau, the maximum occurs when σ_1 is the reverse of σ_2. Later, we shall discuss a variation of Spearman's footrule called "Spearman's rho."

3. Measures for comparing top *k* lists.

We now discuss modifications of these metrics for the case when we have only the top k members of the ordering. Formally, a *top k list* τ is a bijection from a domain D_τ (intuitively, the members of the top k list) to $[k]$. We say that *i appears in* the top k list τ if $i \in D_\tau$. Similar to our convention for permutations, we interpret $\tau(i)$ (for i in D_τ) as the rank of i in τ. As before, we say that *i is ahead of j in τ* if $\tau(i) < \tau(j)$. If τ is a top k list and σ is a permutation on $D \supseteq D_\tau$, then we say that σ is an *extension* of τ, which we denote $\sigma \succeq \tau$, if $\sigma(i) = \tau(i)$ for all $i \in D_\tau$.

Assume that τ_1 and τ_2 are top k lists. In this section, we give several measures for the distance between τ_1 and τ_2. We begin by recalling the definition of a metric and formally defining a distance measure. A binary function d is called *symmetric* if $d(x,y) = d(y,x)$ for all x, y in the domain, and is called *regular* if $d(x,y) = 0$ if and only if $x = y$. We define a *distance measure* to be a nonnegative, symmetric, regular binary function. A *metric* is a distance measure d that satisfies the *triangle inequality* $d(x,z) \leq d(x,y) + d(y,z)$ for all x, y, z in the domain. All of the measures of closeness between top k lists considered in this paper are distance measures.

Global notation. Here we set up some global notation that we use throughout the paper. When two top k lists τ_1 and τ_2 are understood, we write $D = D_{\tau_1} \cup D_{\tau_2}$; $Z = D_{\tau_1} \cap D_{\tau_2}$; $S = D_{\tau_1} \setminus D_{\tau_2}$; $T = D_{\tau_2} \setminus D_{\tau_1}$. Let $z = |Z|$. Note that $|S| = |T| = k - z$ and $|D| = 2k - z$.

Remark. An important feature of our work is that when we compare τ_1 and τ_2, we do not assume that these are top k lists of elements from a fixed domain D. This is a fairly natural requirement in many applications of our work. For example, if we wish to compare the top 10 lists produced by two search engines, it is unreasonable to expect any knowledge of the (possibly very large) universe to which elements of these lists belong; in fact, we cannot even expect to know the size of this universe. The drawback of our requirement is that it is one of the reasons why several very natural distance measures that we define between top k lists fail to be metrics (cf. section 3.3).

3.1. Kendall's tau. There are various natural ways to generalize Kendall's tau to measure distances between top k lists. We now consider some of them. We begin by generalizing the definition of the set $\mathcal{P}$. Given two top k lists τ_1 and τ_2, we define $\mathcal{P}(\tau_1, \tau_2) = \mathcal{P}_{D_{\tau_1} \cup D_{\tau_2}}$ to be the set of all unordered pairs of distinct elements in $D_{\tau_1} \cup D_{\tau_2}$.

For top k lists τ_1 and τ_2, the *minimizing Kendall distance* $K_{\min}(\tau_1, \tau_2)$ between τ_1 and τ_2 is defined to be the minimum value of $K(\sigma_1, \sigma_2)$, where σ_1 and σ_2 are each permutations of $D_{\tau_1} \cup D_{\tau_2}$ and where $\sigma_1 \succeq \tau_1$ and $\sigma_2 \succeq \tau_2$.

For top k lists τ_1 and τ_2, the *averaging Kendall distance* $K_{\text{avg}}(\tau_1, \tau_2)$ between τ_1 and τ_2 is defined to be the expected value $\mathrm{E}(K(\sigma_1, \sigma_2))$, where σ_1 and σ_2 are each permutations of $D_{\tau_1} \cup D_{\tau_2}$ and where $\sigma_1 \succeq \tau_1$ and $\sigma_2 \succeq \tau_2$. Here $\mathrm{E}(\cdot)$ gives the expected value where all extensions are taken to be equally likely.

Next we consider an approach that we will show gives both the minimizing Kendall distance and the averaging Kendall distance as special cases. Let p be a fixed parameter with $0 \leq p \leq 1$. Similar to our definition of $\bar{K}_{i,j}(\sigma_1, \sigma_2)$ for permutations σ_1, σ_2, we define a penalty $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2)$ for top k lists τ_1, τ_2 for $\{i, j\} \in \mathcal{P}(\tau_1, \tau_2)$. There are four cases.

Case 1 (i and j appear in both top k lists). If i and j are in the same order (such as i being ahead of j in both top k lists), then let $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2) = 0$; this corresponds to "no penalty" for $\{i, j\}$. If i and j are in the opposite order (such as i being ahead of j in τ_1 and j being ahead of i in τ_2), then let the penalty $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2) = 1$.

Case 2 (i and j both appear in one top k list (say τ_1), and exactly one of i or j, say i, appears in the other top k list (τ_2)). If i is ahead of j in τ_1, then let the penalty $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2) = 0$, and otherwise let $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2) = 1$. Intuitively, we know that i is ahead of j as far as τ_2 is concerned, since i appears in τ_2 but j does not.

Case 3 (i, but not j, appears in one top k list (say τ_1), and j, but not i, appears in the other top k list (τ_2)). Then let the penalty $\bar{K}^{(p)}_{i,j}(\tau_1, \tau_2) = 1$. Intuitively, we know that i is ahead of j as far as τ_1 is concerned and j is ahead of i as far as τ_2 is concerned.

Case 4 (i and j both appear in one top k list (say τ_1), but neither i nor j appears in the other top k list (τ_2)). This is the interesting case (the only case where there is really an option as to what the penalty should be). We call such pairs $\{i,j\}$ *special pairs.* In this case, we let the penalty $\bar{K}^{(p)}_{i,j}(\tau_1,\tau_2)=p$.

Based on these cases, we now define $K^{(p)}$, the *Kendall distance with penalty parameter* p, as follows:

$$K^{(p)}(\tau_1,\tau_2)=\sum_{\{i,j\}\in\mathcal{P}(\tau_1,\tau_2)}\bar{K}^{(p)}_{i,j}(\tau_1,\tau_2).$$

When $p=0$, this gives an "optimistic approach." It corresponds to the intuition that we assign a nonzero penalty score to the pair $\{i,j\}$ only if we have enough information to know that i and j are in the opposite order according to the two top k lists. When $p=1/2$, this gives a "neutral approach." It corresponds to the intuition that we do not have enough information to know whether the penalty score should be 0 or 1, so we assign a neutral penalty score of 1/2. Later, we show that the optimistic approach gives precisely $K_{\min}$ and the neutral approach gives precisely K_{avg}.

The next lemma gives a formula, which we shall find useful later, for $K^{(p)}$.

LEMMA 3.1. $K^{(p)}(\tau_1,\tau_2)=(k-z)((2+p)k-pz+1-p)+\sum_{i,j\in Z}\bar{K}^{(0)}_{i,j}(\tau_1,\tau_2)-\sum_{j\in S}\tau_1(j)-\sum_{j\in T}\tau_2(j)$.

Proof. We analyze the four cases in the definition of $K^{(p)}(\tau_1,\tau_2)$ and obtain formulas for each of them in terms of our global notation. Case 1 is the situation when for a pair $\{i,j\}$, we have $i,j\in Z$. In this case, the contribution of this pair to $K^{(p)}(\tau_1,\tau_2)$ is

$$\sum_{i,j\in Z}\bar{K}^{(0)}_{i,j}(\tau_1,\tau_2). \tag{1}$$

Case 2 is the situation when for a pair $\{i,j\}$, one of i or j is in Z and the other is in either S or T. Let us denote by i the element in Z and by j the element in S or T. Let us now consider the case when $i\in Z, j\in S$. Let $j_1<\cdots<j_{k-z}$ be the elements in S. Fix an $\ell\in\{1,\ldots,k-z\}$ and consider the element j_ℓ and its rank $\tau_1(j_\ell)$ in the first top k list τ_1. There will be a contribution of 1 to $K^{(p)}(\tau_1,\tau_2)$ for all $i\in Z$ such that $\tau_1(i)>\tau_1(j_\ell)$, that is, all the elements $i\in Z$ such that j_ℓ is ahead of i in τ_1; denote this net contribution of ℓ to $K^{(p)}(\tau_1,\tau_2)$ by $\gamma(\ell)$. We now obtain an expression for $\gamma(\ell)$. The total number of elements that j_ℓ is ahead of in τ_1 is $k-\tau_1(j_\ell)$, and of these elements, $\ell-1$ of them belong to S and the rest belong to Z. This gives $\gamma(\ell)=k-\tau_1(j_\ell)-(\ell-1)$. Now, summing over all ℓ, the contribution to $K^{(p)}(\tau_1,\tau_2)$ is $\sum_{\ell=1}^{k-z}\gamma(\ell)=(k-z)(k+z+1)/2-\sum_{j\in S}\tau_1(j)$. Similarly, for the case when $i\in Z, j\in T$, the contribution to $K^{(p)}(\tau_1,\tau_2)$ is $(k-z)(k+z+1)/2-\sum_{j\in T}\tau_2(j)$. Summing these, the term corresponding to Case 2 contributing to $K^{(p)}(\tau_1,\tau_2)$ is

$$(k-z)(k+z+1)-\sum_{j\in S}\tau_1(j)-\sum_{j\in T}\tau_2(j). \tag{2}$$

Case 3 is the situation when for a pair $\{i,j\}$, we have $i\in S$ and $j\in T$. The total contribution to $K^{(p)}(\tau_1,\tau_2)$ from this case is

$$|S|\times|T|=(k-z)^2. \tag{3}$$

Finally, Case 4 is the situation when for a pair $\{i, j\}$, we have either $i, j \in S$ or $i, j \in T$. The total contribution to $K^{(p)}(\tau_1, \tau_2)$ from this case is

$$p\binom{|S|}{2} + p\binom{|T|}{2} = 2p\binom{k-z}{2}. \tag{4}$$

Adding equations (1)–(4), we obtain

$$K^{(p)}(\tau_1, \tau_2) = (k-z)((2+p)k-pz+1-p)+\sum_{i,j\in Z} \bar{K}^{(0)}_{i,j}(\tau_1, \tau_2)-\sum_{j\in S}\tau_1(j)-\sum_{j\in T}\tau_2(j). \qquad \square$$

Let A and B be finite sets of objects (in our case of interest, these objects are permutations). Let d be a metric of distances between objects (at the moment, we are interested in the case where d is the Kendall distance between permutations). The *Hausdorff distance* between A and B is given by

$$d_{\text{Haus}}(A, B) = \max\left\{\max_{\sigma_1\in A}\min_{\sigma_2\in B} d(\sigma_1, \sigma_2), \max_{\sigma_2\in B}\min_{\sigma_1\in A} d(\sigma_1, \sigma_2)\right\}.$$

Although this looks fairly nonintuitive, it is actually quite natural, as we now explain. The quantity $\min_{\sigma_2\in B} d(\sigma_1, \sigma_2)$ is the distance between σ_1 and the set B. Therefore, the quantity $\max_{\sigma_1\in A}\min_{\sigma_2\in B} d(\sigma_1, \sigma_2)$ is the maximal distance of a member of A from the set B. Similarly, the quantity $\max_{\sigma_2\in B}\min_{\sigma_1\in A} d(\sigma_1, \sigma_2)$ is the maximal distance of a member of B from the set A. Therefore, the Hausdorff distance between A and B is the maximal distance of a member of A or B from the other set. Thus, A and B are within Hausdorff distance s of each other precisely if every member of A and B is within distance s of some member of the other set. The Hausdorff distance is well known to be a metric.

Critchlow [Cri80] used the Hausdorff distance to define a distance measure between top k lists. Specifically, given a metric d that gives the distance between permutations, Critchlow defined the distance between top k lists τ_1 and τ_2 to be

$$\max\left\{\max_{\sigma_1\succeq\tau_1}\min_{\sigma_2\succeq\tau_2} d(\sigma_1, \sigma_2), \max_{\sigma_2\succeq\tau_2}\min_{\sigma_1\succeq\tau_1} d(\sigma_1, \sigma_2)\right\}. \tag{5}$$

Critchlow assumed that there is a fixed domain D, and so σ_1 and σ_2 range over all permutations with domain D. This distance measure is a metric, since it is a special case of a Hausdorff metric.

We, too, are interested in considering a version of the Hausdorff distance. However, as remarked earlier, in this paper we do not assume a fixed domain. Therefore, we define K_{Haus}, the Hausdorff version of the Kendall distance between top k lists, to be given by (5) with $d(\sigma_1, \sigma_2)$ as the Kendall distance $K(\sigma_1, \sigma_2)$, but where, unlike Critchlow, we take σ_1 and σ_2 to be permutations of $D_{\tau_1} \cup D_{\tau_2}$.

Critchlow obtains a closed form for his version of (5) when $d(\sigma_1, \sigma_2)$ is the Kendall distance $K(\sigma_1, \sigma_2)$. Specifically, if n is the size of the underlying domain D, and $d(\sigma_1, \sigma_2) = K(\sigma_1, \sigma_2)$, he shows that (5) is given by

$$(k-z)\left(n+k-\frac{k-z-1}{2}\right)+\sum_{i,j\in Z} \bar{K}^{(0)}_{i,j}(\tau_1, \tau_2)-\sum_{i\in S}\tau_1(i)-\sum_{i\in T}\tau_2(i). \tag{6}$$

By replacing n by $2k-z$, we obtain a closed form for K_{Haus}.

LEMMA 3.2.

$$K_{\mathrm{Haus}}(\tau_1,\tau_2)=\frac{1}{2}(k-z)(5k-z+1)+\sum_{i,j\in Z}\bar{K}^{(0)}_{i,j}(\tau_1,\tau_2)-\sum_{i\in S}\tau_1(i)-\sum_{i\in T}\tau_2(i).$$

We show that the "optimistic approach" given by $K^{(0)}$ and the "neutral approach" given by $K^{(1/2)}$ are exactly $K_{\min}$ and K_{avg}, respectively. Furthermore, we show the somewhat surprising result that the Hausdorff distance K_{Haus} also equals $K^{(1/2)}$.

PROPOSITION 3.3. $K_{\min}=K^{(0)}$.

Proof. Let τ_1 and τ_2 be top k lists. We must show that $K_{\min}(\tau_1,\tau_2)=K^{(0)}(\tau_1,\tau_2)$. Define σ_1 to be the extension of τ_1 over D where the elements are, in order, the elements of D_{τ_1} in the same order as they are in τ_1, followed by the elements of T in the same order as they are in τ_2. For example, if $k=4$, if the top 4 elements of τ_1 are, in order, 1, 2, 3, 4, and if the top 4 elements of τ_2 are, in order, 5, 4, 2, 6, then the ordering of the elements for σ_1 is 1, 2, 3, 4, 5, 6. We similarly define the extension σ_2 of τ_2 by reversing the roles of τ_1 and τ_2. First, we show that $K_{\min}(\tau_1,\tau_2)=K(\sigma_1,\sigma_2)$, and then we show that $K(\sigma_1,\sigma_2)=K^{(0)}(\tau_1,\tau_2)$.

To show that $K_{\min}(\tau_1,\tau_2)=K(\sigma_1,\sigma_2)$, it is clearly sufficient to show that if σ_1' is an arbitrary extension of τ_1 (over D) and σ_2' is an arbitrary extension of τ_2 (over D), and if $\{i,j\}$ is an arbitrary member of $\mathcal{P}(\tau_1,\tau_2)$, then

$$\bar{K}_{i,j}(\sigma_1,\sigma_2)\le\bar{K}_{i,j}(\sigma_1',\sigma_2'). \tag{7}$$

When $\{i,j\}$ is not a special pair (that is, when $\{i,j\}$ falls into the first three cases of the definition of $\bar{K}^{(p)}_{i,j}(\tau_1,\tau_2)$), we have equality in (7), since the ordering of i and j according to $\sigma_1,\sigma_2,\sigma_1',\sigma_2'$ are forced by τ_1,τ_2. When $\{i,j\}$ is a special pair, we have $\bar{K}_{i,j}(\sigma_1,\sigma_2)=0$, and so again (7) holds.

We have shown that $K_{\min}(\tau_1,\tau_2)=K(\sigma_1,\sigma_2)$. Hence, we need only show that $K^{(0)}(\tau_1,\tau_2)=K(\sigma_1,\sigma_2)$. To show this, we need only show that $\bar{K}^{(0)}_{i,j}(\tau_1,\tau_2)=\bar{K}_{i,j}(\sigma_1,\sigma_2)$ for every pair $\{i,j\}$. As before, this is automatic when $\{i,j\}$ is not a special pair. When $\{i,j\}$ is a special pair, we have $\bar{K}^{(0)}_{i,j}(\tau_1,\tau_2)=0=\bar{K}_{i,j}(\sigma_1,\sigma_2)$. This concludes the proof. □

PROPOSITION 3.4. $K_{\mathrm{avg}}=K^{(1/2)}=K_{\mathrm{Haus}}$.

Proof. Let τ_1,τ_2 be top k lists. Then

$$\begin{aligned}K_{\mathrm{avg}}(\tau_1,\tau_2)&=\mathrm{E}(K(\sigma_1,\sigma_2))\\&=\mathrm{E}\left(\sum_{\{i,j\}\in\mathcal{P}(\tau_1,\tau_2)}\bar{K}_{i,j}(\sigma_1,\sigma_2)\right)\\&=\sum_{\{i,j\}\in\mathcal{P}(\tau_1,\tau_2)}\mathrm{E}\left(\bar{K}_{i,j}(\sigma_1,\sigma_2)\right). \qquad (8)\end{aligned}$$

We shall show that

$$\mathrm{E}\left(\bar{K}_{i,j}(\sigma_1,\sigma_2)\right)=\bar{K}^{(1/2)}_{i,j}(\tau_1,\tau_2). \tag{9}$$

This proves that $K_{\mathrm{avg}}=K^{(1/2)}$, since the result of substituting $\bar{K}^{(1/2)}_{i,j}(\tau_1,\tau_2)$ for $\mathrm{E}(\bar{K}_{i,j}(\sigma_1,\sigma_2))$ in (8) gives $K^{(1/2)}(\tau_1,\tau_2)$. Similar to before, when $\{i,j\}$ is not a special pair, we have $\bar{K}_{i,j}(\sigma_1,\sigma_2)=\bar{K}^{(1/2)}(\tau_1,\tau_2)$, and so (9) holds. When $\{i,j\}$ is a special pair, then $\bar{K}_{i,j}^{(1/2)}(\tau_1, \tau_2) = 1/2$. So we are done with showing that $K_{\text{avg}} = K^{(1/2)}$ if we show that when $\{i, j\}$ is a special pair, then $\mathrm{E}(\bar{K}_{i,j}(\sigma_1, \sigma_2)) = 1/2$. Assume without loss of generality that i, j are both in D_{τ_1} but neither is in D_{τ_2}. The ordering of i, j in σ_1 is forced by τ_1. Further, there is a one-to-one correspondence between those permutations σ_2 that extend τ_2 with i ahead of j and those that extend τ_2 with j ahead of i (the correspondence is determined by simply switching i and j). Therefore, for each choice of σ_1, exactly half of the choices for σ_2 have $\bar{K}_{i,j}(\sigma_1, \sigma_2) = 0$, and for the other half, $\bar{K}_{i,j}(\sigma_1, \sigma_2) = 1$. So $\mathrm{E}(\bar{K}_{i,j}(\sigma_1, \sigma_2)) = 1/2$, as desired.

We now show that $K_{\text{Haus}} = K^{(1/2)}$. If we set $p = 1/2$ in our formula for $K^{(p)}$ given in Lemma 3.1, we obtain the right-hand side of the equation in Lemma 3.2. Thus, $K_{\text{Haus}} = K^{(1/2)}$. We now give a direct proof that does not require the use of Lemma 3.2 and hence does not require the use of Critchlow's formula given by (6).

Let τ_1, τ_2 be top k lists. Then $K_{\text{Haus}}(\tau_1, \tau_2)$ is given by

$$\max\left\{\max_{\sigma_1 \succeq \tau_1} \min_{\sigma_2 \succeq \tau_2} K(\sigma_1, \sigma_2), \max_{\sigma_2 \succeq \tau_2} \min_{\sigma_1 \succeq \tau_1} K(\sigma_1, \sigma_2)\right\}.$$

Let σ_1^* be the permutation over $D_{\tau_1} \cup D_{\tau_2}$ where $\sigma_1^* \succeq \tau_1$ and where $\sigma_1^*(k+1), \ldots, \sigma_1^*(2k - z)$ are, respectively, the members of T in reverse order. Let σ_2^* be the permutation over $D_{\tau_1} \cup D_{\tau_2}$ where $\sigma_2^* \succeq \tau_2$ and where $\sigma_2^*(k+1), \ldots, \sigma_2^*(2k - z)$ are, respectively, the members of S in order (not in reverse order). It is not hard to see that $K_{\text{Haus}}(\tau_1, \tau_2) = K(\sigma_1^*, \sigma_2^*)$. So we need only show that $K(\sigma_1^*, \sigma_2^*) = K^{(1/2)}(\tau_1, \tau_2)$.

In the definition of $K^{(p)}$, let us consider the contribution of each pair $\{i, j\}$ to $K^{(1/2)}(\tau_1, \tau_2)$, as compared to its contribution to $K(\sigma_1^*, \sigma_2^*)$. In the first three cases in the definition of $K^{(p)}$, it is easy to see that $\{i, j\}$ contributes exactly the same to $K^{(1/2)}(\tau_1, \tau_2)$ as to $K(\sigma_1^*, \sigma_2^*)$. Let us now consider Case 4, where $\{i, j\}$ is a special pair, that is, where both i and j appear in one of the top k lists τ_1 or τ_2, but neither appears in the other top k list. If both i and j appear in τ_1 but neither appears in τ_2, then the contribution to $K^{(1/2)}(\tau_1, \tau_2)$ is $1/2$, and the contribution to $K(\sigma_1^*, \sigma_2^*)$ is 0. If both i and j appear in τ_2 but neither appears in τ_1, then the contribution to $K^{(1/2)}(\tau_1, \tau_2)$ is $1/2$ and the contribution to $K(\sigma_1^*, \sigma_2^*)$ is 1. Since there are just as many pairs $\{i, j\}$ of the first type (where both i and j appear in τ_1 but neither appears in τ_2) as there are of the second type (where both i and j appear in τ_2 but neither appears in τ_1), the total contribution of all pairs $\{i, j\}$ of Case 4 to $K^{(1/2)}(\tau_1, \tau_2)$ and $K(\sigma_1^*, \sigma_2^*)$ is the same. This proves that $K_{\text{Haus}} = K^{(1/2)}$. □

3.2. Spearman's footrule. We now generalize Spearman's footrule to several methods for determining distances between top k lists, just as we did for Kendall's tau.

For top k lists τ_1 and τ_2, the *minimizing footrule distance* $F_{\min}(\tau_1, \tau_2)$ between τ_1 and τ_2 is defined to be the minimum value of $F(\sigma_1, \sigma_2)$, where σ_1 and σ_2 are each permutations of D and where $\sigma_1 \succeq \tau_1$ and $\sigma_2 \succeq \tau_2$.

For top k lists τ_1 and τ_2, the *averaging footrule distance* $F_{\text{avg}}(\tau_1, \tau_2)$ between τ_1 and τ_2 is defined to be the expected value $\mathrm{E}(F(\sigma_1, \sigma_2))$, where σ_1 and σ_2 are each permutations of $D_{\tau_1} \cup D_{\tau_2}$ and where $\sigma_1 \succeq \tau_1$ and $\sigma_2 \succeq \tau_2$. Again, $\mathrm{E}(\cdot)$ gives the expected value where all extensions are taken to be equally likely.

Let ℓ be a real number greater than k. The *footrule distance with location parameter* ℓ, denoted $F^{(\ell)}$, is obtained—intuitively—by placing all missing elements in each of the lists at position ℓ and computing the usual footrule distance between them. More formally, given top k lists τ_1 and τ_2, define functions τ_1' and τ_2' with domain

$D_{\tau_1} \cup D_{\tau_2}$ by letting $\tau_1'(i) = \tau_1(i)$ for $i \in D_{\tau_1}$ and $\tau_1'(i) = \ell$ otherwise, and similarly defining τ_2'. We then define $F^{(\ell)}$ by setting $F^{(\ell)}(\tau_1, \tau_2) = \sum_{i \in D_{\tau_1} \cup D_{\tau_2}} |\tau_1'(i) - \tau_2'(i)|$.

A natural choice for ℓ is $k+1$, and we make this choice in our experiments (section 9). We denote $F^{(k+1)}$ simply by F^*.

The next lemma gives a formula, which we shall find useful later, for $F^{(\ell)}$.

LEMMA 3.5. $F^{(\ell)}(\tau_1, \tau_2) = 2(k-z)\ell + \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| - \sum_{i \in S} \tau_1(i) - \sum_{i \in T} \tau_2(i)$.

Proof.

$$\begin{aligned} F^{(\ell)}(\tau_1, \tau_2) &= \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| + \sum_{i \in S} |\tau_1(i) - \tau_2(i)| + \sum_{i \in T} |\tau_1(i) - \tau_2(i)| \\ &= \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| + \sum_{i \in S} (\ell - \tau_1(i)) + \sum_{i \in T} (\ell - \tau_2(i)) \\ &= 2(k-z)\ell + \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| - \sum_{i \in S} \tau_1(i) - \sum_{i \in T} \tau_2(i). \qquad \square \end{aligned}$$

Similar to our definition of K_{Haus}, we define F_{Haus}, the Hausdorff version of the footrule distance between top k lists, to be given by (5) with $d(\sigma_1, \sigma_2)$ as the footrule distance $F(\sigma_1, \sigma_2)$, where, as before, we take σ_1 and σ_2 to be permutations of $D_{\tau_1} \cup D_{\tau_2}$.

Just as he did with the Kendall distance, Critchlow considered his version of (5) when $d(\sigma_1, \sigma_2)$ is the footrule distance $F(\sigma_1, \sigma_2)$ and where there is a fixed domain of size n. He obtained a closed formula given by

$$(k-z)(2n+1-(k-z)) + \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| - \sum_{i \in S} \tau_1(i) - \sum_{i \in T} \tau_2(i).$$

By replacing n by $2k - z$, we obtain a closed form for F_{Haus}.

LEMMA 3.6.

$$\begin{aligned} F_{\text{Haus}}(\tau_1, \tau_2) &= (k-z)(3k-z+1) + \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| - \sum_{i \in S} \tau_1(i) - \sum_{i \in T} \tau_2(i) \\ &= F^{(\frac{3k-z+1}{2})}(\tau_1, \tau_2). \end{aligned}$$

The last equality is obtained by formally substituting $\ell = (3k - z + 1)/2$ into the formula for $F^{(\ell)}$ given by Lemma 3.5. Thus, intuitively, $F_{\text{Haus}}(\tau_1, \tau_2)$ is a "dynamic" version of $F^{(\ell)}$, where $\ell = (3k - z + 1)/2$ actually depends on τ_1 and τ_2. Since $F_{\min} = F_{\text{avg}} = F_{\text{Haus}}$ (Proposition 3.7), this gives us a formula for $F_{\min}$ and F_{avg} as well. Note that $\ell = (3k - z + 1)/2$ is the average of $k+1$ and $2k - z$, where the latter number is the size of $D = D_{\tau_1} \cup D_{\tau_2}$. Since taking $\ell = (3k - z + 1)/2$ corresponds intuitively to "placing the missing elements at an average location," it is not surprising that the resulting formula gives F_{avg}.

Unlike the situation with $K_{\min}$ and K_{avg}, the next proposition tells us that $F_{\min}$ and F_{avg} are the same. Furthermore, the Hausdorff distance F_{Haus} shares this common value.

PROPOSITION 3.7. $F_{\min} = F_{\text{avg}} = F_{\text{Haus}}$.

Proof. Let τ_1 and τ_2 be top k lists. Let $\sigma_1, \sigma_1', \sigma_2, \sigma_2'$ be permutations of $D = D_{\tau_1} \cup D_{\tau_2}$, where σ_1 and σ_1' extend τ_1 and where σ_2 and σ_2' extend τ_2. We need only show that $F(\sigma_1, \sigma_2) = F(\sigma_1', \sigma_2')$, that is, that the value of $F(\sigma_1, \sigma_2)$ is independent of the choice of σ_1, σ_2. Therefore, we need only show that $F(\sigma_1, \sigma_2) = F(\sigma_1, \sigma_2')$, where σ_1 is held fixed, since by symmetry (where σ_2' is held fixed) we would then have $F(\sigma_1, \sigma_2') = F(\sigma_1', \sigma_2')$, and hence $F(\sigma_1, \sigma_2) = F(\sigma_1, \sigma_2') = F(\sigma_1', \sigma_2')$, as desired.

Now $F(\sigma_1, \sigma_2) = \sum_{i \in D} |\sigma_1(i) - \sigma_2(i)|$. So we need only show that

$$\sum_{i \in D} |\sigma_1(i) - \sigma_2(i)| = \sum_{i \in D} |\sigma_1(i) - \sigma_2'(i)|. \tag{10}$$

Now

$$\sum_{i \in D} |\sigma_1(i) - \sigma_2(i)| = \sum_{i \in D_{\tau_2}} |\sigma_1(i) - \sigma_2(i)| + \sum_{i \in S} |\sigma_1(i) - \sigma_2(i)|, \tag{11}$$

and similarly

$$\sum_{i \in D} |\sigma_1(i) - \sigma_2'(i)| = \sum_{i \in D_{\tau_2}} |\sigma_1(i) - \sigma_2'(i)| + \sum_{i \in S} |\sigma_1(i) - \sigma_2'(i)|. \tag{12}$$

Now $\sigma_2(i) = \sigma_2'(i)$ for $i \in D_{\tau_2}$. Hence,

$$\sum_{i \in D_{\tau_2}} |\sigma_1(i) - \sigma_2(i)| = \sum_{i \in D_{\tau_2}} |\sigma_1(i) - \sigma_2'(i)|. \tag{13}$$

From (11), (12), and (13), it follows that to prove (10), and hence complete the proof, it is sufficient to prove

$$\sum_{i \in S} |\sigma_1(i) - \sigma_2(i)| = \sum_{i \in S} |\sigma_1(i) - \sigma_2'(i)|. \tag{14}$$

If $i \in S$, then $\sigma_1(i) \leq k < \sigma_2(i)$. Thus, if $i \in S$, then $\sigma_1(i) < \sigma_2(i)$, and similarly $\sigma_1(i) < \sigma_2'(i)$. So it is sufficient to prove

$$\sum_{i \in S} (\sigma_1(i) - \sigma_2(i)) = \sum_{i \in S} (\sigma_1(i) - \sigma_2'(i))$$

and hence to prove

$$\sum_{i \in S} \sigma_2(i) = \sum_{i \in S} \sigma_2'(i). \tag{15}$$

But both the left-hand side and the right-hand side of (15) equal $\sum_{\ell=k+1}^{|D|} \ell$, and hence are equal. This completes the proof that $F_{\min} = F_{\text{avg}} = F_{\text{Haus}}$. □

3.3. Metric properties. We have now introduced three distinct measures of closeness between top k lists: (1) $K^{(p)}$, which has $K_{\min}$ and $K_{\text{avg}} = K_{\text{Haus}}$ as special cases for certain choices of p; (2) $F_{\min}$, which equals F_{avg} and F_{Haus}; and (3) $F^{(\ell)}$. Perhaps the most natural question, and the main subject of our investigation, is to ask whether or not they are metrics.

As a preview to our main results, we begin by observing that while $F^{(\ell)}$ is a metric, none of the other distance measures that we have defined (namely, $K^{(p)}$ and $F_{\min}$, hence also $K_{\min}, K_{\text{avg}}, K_{\text{Haus}}, F_{\text{avg}}, F_{\text{Haus}}$) is a metric.

Proposition 3.8. *The distance measure $F^{(\ell)}$ is a metric for every choice of the location parameter ℓ.*

Proof. We need only show that the triangle inequality holds. Let τ_1, τ_2, τ_3 be top k lists. Let $n = |D_{\tau_1} \cup D_{\tau_2} \cup D_{\tau_3}|$. Define an n-dimensional vector v_1 corresponding to τ_1 by letting $v_1(i) = \tau_1(i)$ for $i \in D_{\tau_1}$ and ℓ otherwise. Similarly, define an n-dimensional vector v_2 corresponding to τ_2 and an n-dimensional vector v_3 corresponding to τ_3. It is easy to see that $F^{(\ell)}(\tau_1, \tau_2)$ is the L_1 distance between v_1 and v_2 and similarly for $F^{(\ell)}(\tau_1, \tau_3)$ and $F^{(\ell)}(\tau_2, \tau_3)$. The triangle inequality for $F^{(\ell)}$ then follows immediately from the triangle inequality for the L_1 norm between two vectors in n-dimensional Euclidean space. □

The other two distinct distance measures, namely $K^{(p)}$ and $F_{\min}$, are not metrics, as we now show. Let τ_1 be the top 2 list where the top 2 items in order are 1,2; let τ_2 be the top 2 list where the top 2 items in order are 1,3; let τ_3 be the top 2 list where the top 2 items in order are 3, 4. It is straightforward to verify that $K^{(p)}(\tau_1, \tau_2) = 1$, $K^{(p)}(\tau_1, \tau_3) = 4 + 2p$, and $K^{(p)}(\tau_2, \tau_3) = 2$. So the triangle inequality fails, because $K^{(p)}(\tau_1, \tau_3) > K^{(p)}(\tau_1, \tau_2) + K^{(p)}(\tau_2, \tau_3)$ for every $p \geq 0$. Therefore, $K^{(p)}$ is not a metric, no matter what the choice of the penalty parameter p is; in particular, by Propositions 3.3 and 3.4, neither $K_{\min}$ nor K_{avg} is a metric.

The same counterexample shows that $F_{\min}$ is not a metric. In this case, it is easy to verify that $F_{\min}(\tau_1, \tau_2) = 2$, $F_{\min}(\tau_1, \tau_3) = 8$, and $F_{\min}(\tau_2, \tau_3) = 4$. So the triangle inequality fails, because $F_{\min}(\tau_1, \tau_3) > F_{\min}(\tau_1, \tau_2) + F_{\min}(\tau_2, \tau_3)$.

The fact that $F_{\min}$ (and hence F_{avg} and F_{Haus}) are not metrics shows that they are not special cases of $F^{(\ell)}$, since $F^{(\ell)}$ is a metric. This is in contrast to the situation with Kendall distances, where $K_{\min}$, K_{avg}, and K_{Haus} are special cases of $K^{(p)}$. (As we noted earlier, the versions of F_{Haus} and K_{Haus} defined by Critchlow [Cri80] are indeed metrics, since the domain is fixed in his case.)

4. Metrics, near metrics, and equivalence classes.

Motivated by the fact that most of our distance measures are not metrics (except for the somewhat strange measure $F^{(\ell)}$), we next consider a precise sense in which each is a "near metric." Actually, we shall consider two seemingly different notions of being a near metric, which our distance measures satisfy, and obtain the surprising result that these notions are actually equivalent.

Our first notion of near metric is based on "relaxing" the triangle inequality (or more generally, the polygonal inequality) that a metric is supposed to satisfy.

Definition 4.1 (relaxed inequalities). *A binary function d satisfies the* c-triangle inequality *if $d(x, z) \leq c(d(x, y) + d(y, z))$ for all x, y, z in the domain. A binary function d satisfies the* c-polygonal inequality *if $d(x, z) \leq c(d(x, x_1) + d(x_1, x_2) + \cdots + d(x_{n-1}, z))$ for all $n > 1$ and $x, z, x_1, \ldots, x_{n-1}$ in the domain.*

The notion of c-triangle inequality, to our knowledge, appears to be rarely studied. It has been used in a paper on pattern matching [FS98] and in the context of the traveling salesperson problem [AB95, BC00]. We do not know if the c-polygonal inequality has ever been studied.

Definition 4.2 (relaxed metrics). *A* c-relaxed$_{\text{t}}$ metric *is a distance measure that satisfies the c-triangle inequality. A* c-relaxed$_{\text{p}}$ metric *is a distance measure that satisfies the c-polygonal inequality.*

Of course, every c-relaxed$_{\text{p}}$ metric is a c-relaxed$_{\text{t}}$ metric. Theorem 4.7 below says that there is a c-relaxed$_{\text{t}}$ metric that is not a c'-relaxed$_{\text{p}}$ metric for any constant c'. We shall focus here on the stronger notion of being a c-relaxed$_{\text{p}}$ metric.

The other notion of near metric that we now discuss is based on bounding the distance measure above and below by positive constant multiples of a metric.

DEFINITION 4.3 (metric boundedness). *A* (c_1, c_2)-metric-bounded distance measure *is a distance measure d for which there is a metric d' and positive constants c_1 and c_2 such that $c_1 d'(x,y) \leq d(x,y) \leq c_2 d'(x,y)$.*

Note that without loss of generality, we can take $c_1 = 1$ (by replacing the metric d' by the metric $c_1 d'$). In this case, we say that d is *c_2-metric bounded.*

The next theorem gives the unexpected result that our two notions of near metric are equivalent (and even with the same value of c).

THEOREM 4.4 (main result 1). *Let d be a distance measure. Then d is a c-relaxed$_{\mathrm{p}}$ metric if and only if d is c-metric-bounded.*

Proof. $\Longleftarrow$ Assume that d is a c-relaxed$_{\mathrm{p}}$ metric. Define d' by taking

$$d'(x,z) = \min_{\ell} \min_{y_0,\ldots,y_\ell \mid y_0 = x \text{ and } y_\ell = z} \sum_{i=0}^{\ell-1} d(y_i, y_{i+1}). \tag{16}$$

We now show that d' is a metric.

First, we have $d'(x,x) = 0$, since $d(x,x) = 0$. From (16) and the polygonal inequality with constant c, we have $d'(x,z) \geq (1/c)d(x,z)$. Hence, $d'(x,z) \neq 0$ if $x \neq z$. Symmetry of d' follows immediately from symmetry of d. Finally, d' satisfies the triangle inequality, since

$$\begin{aligned} d'(x,z) &= \min_{\ell} \min_{y_0,\ldots,y_\ell \mid y_0 = x \text{ and } y_\ell = z} \sum_{i=0}^{\ell-1} d(x_i, x_{i+1}) \\ &\leq \min_{\ell_1} \min_{y_0,\ldots,y_{\ell_1} \mid y_0 = x \text{ and } y_{\ell_1} = y} \sum_{i=0}^{\ell_1-1} d(y_i, y_{i+1}) \\ &\quad + \min_{\ell_2} \min_{z_0,\ldots,z_{\ell_1} \mid z_0 = y \text{ and } z_{\ell_2} = z} \sum_{i=0}^{\ell_2-1} d(z_i, z_{i+1}) \\ &= d'(x,y) + d'(y,z). \end{aligned}$$

Therefore, d' is a metric.

We now show that d is c-metric-bounded. By (16), it follows easily that $d'(x,z) \leq d(x,z)$. By (16) and the polygonal inequality with constant c, we have $d(x,z) \leq cd'(x,z)$.

$\Longrightarrow$ Assume that d is c-metric-bounded. Then $0 = d'(x,x) \leq d(x,x) \leq cd'(x,x) = 0$. Therefore, $d(x,x) = 0$. If $x \neq y$, then $d(x,y) \geq d'(x,y) > 0$. We now show that d satisfies the c-polygonal inequality.

$$\begin{aligned} d(x,z) &\leq cd'(x,z) \\ &\leq c(d'(x,x_1) + d'(x_1,x_2) + \cdots + d'(x_{n-1},z)) \text{ since } d' \text{ is a metric} \\ &\leq c(d(x,x_1) + d(x_1,x_2) + \cdots + d(x_{n-1},z)) \text{ since } d'(x,y) \leq d(x,y). \end{aligned}$$

Since also d is symmetric by assumption, it follows that d is a c-relaxed$_{\mathrm{p}}$ metric. □

Inspired by Theorem 4.4, we now define what it means for a distance measure to be "almost" a metric, and a robust notion of "similar" or "equivalent" distance measures.

DEFINITION 4.5 (near metric). *A distance measure between top k lists is a* near metric *if there is a constant c, independent of k, such that the distance measure is a c-relaxed$_{\mathrm{p}}$ metric (or, equivalently, is c-metric-bounded).*[2]

DEFINITION 4.6 (equivalent distance measures). *Two distance measures d and d' between top k lists are* equivalent *if there are positive constants c_1 and c_2 such that $c_1 d'(\tau_1, \tau_2) \le d(\tau_1, \tau_2) \le c_2 d'(\tau_1, \tau_2)$ for every pair τ_1, τ_2 of top k lists.*[3]

It is easy to see that this definition of equivalence actually gives us an equivalence relation (reflexive, symmetric, and transitive). It follows from Theorem 4.4 that a distance measure is equivalent to a metric if and only if it is a near metric.

Our notion of equivalence is inspired by a classical result of Diaconis and Graham [DG77], which states that for every two permutations σ_1, σ_2, we have

$$K(\sigma_1, \sigma_2) \le F(\sigma_1, \sigma_2) \le 2K(\sigma_1, \sigma_2). \tag{17}$$

(Of course, we are dealing with distances between top k lists, whereas Diaconis and Graham dealt with distances between permutations.)

Having showed that the notions of c-relaxed$_{\mathrm{p}}$ metric and c-metric-boundedness are identical, we compare these to the notions of c-relaxed$_{\mathrm{t}}$ metric and the classical topological notion of being a topological metric, that is, of generating a metrizable topology.

THEOREM 4.7. *Every c-relaxed$_{\mathrm{p}}$ metric is a c-relaxed$_{\mathrm{t}}$ metric, but not conversely. In fact, there is a c-relaxed$_{\mathrm{t}}$ metric that is not a c'-relaxed$_{\mathrm{p}}$ metric for any constant c'.*

Proof. It is clear that every c-relaxed$_{\mathrm{p}}$ metric is a c-relaxed$_{\mathrm{t}}$ metric. We now show that the converse fails. Define d on the space $[0, 1]$ by taking $d(x, y) = (x - y)^2$. It is clear that d is a symmetric function with $d(x, y) = 0$ if and only if $x = y$. To show the 2-triangle inequality, let $\alpha = d(x, z)$, $\beta = d(x, y)$, and $\gamma = d(y, z)$. Now $\sqrt{\alpha} \le \sqrt{\beta} + \sqrt{\gamma}$, since the function d' with $d'(x, y) = |x - y|$ is a metric. By squaring both sides, we get $\alpha \le \beta + \gamma + 2\sqrt{\beta\gamma}$. But $\sqrt{\beta\gamma} \le (\beta + \gamma)/2$ by the well-known fact that the geometric mean is bounded above by the arithmetic mean. We therefore obtain $\alpha \le 2(\beta + \gamma)$, that is, $d(x, z) \le 2(d(x, y) + d(y, z))$. So d is a 2-relaxed$_{\mathrm{t}}$ metric.

Let n be an arbitrary positive integer, and define x_i to be i/n for $1 \le i \le n - 1$. Then $d(0, x_1) + d(x_1, x_2) + \cdots + d(x_{n-1}, 1) = n(1/n^2) = 1/n$. Since this converges to 0 as n goes to infinity, and since $d(0, 1) = 1$, there is no constant c' for which d satisfies the polygonal inequality. Therefore, d is a c-relaxed$_{\mathrm{t}}$ metric that is not a c'-relaxed$_{\mathrm{p}}$ metric for any constant c'. □

THEOREM 4.8. *Every c-relaxed$_{\mathrm{t}}$ metric is a topological metric, but not conversely. The converse fails even if we restrict attention to distance measures.*

Proof. By the *topological space induced by a binary function d*, we mean the topological space whose open sets are precisely the union of sets ("ϵ-balls") of the form $\{y \mid d(x, y) < \epsilon\}$. A topological space is *metrizable* if there is a metric d that induces the topology. A *topological metric* is a binary function d such that the topology induced by d is metrizable.

There is a theorem of Nagata and Smirnov [Dug66, pp. 193–195] that a topological space is metrizable if and only if it is regular and has a basis that can be decomposed

[2] It makes sense to say that the constant c is independent of k, since each of our distance measures is actually a family, parameterized by k. We need to make an assumption that c is independent of k, since otherwise we are simply considering distance measures over finite domains, where there is always such a constant c.

[3] As before, the constants c_1 and c_2 are assumed to be independent of k.

into an at most countable collection of neighborhood-finite families. The proof of the "only if" direction can be modified in an obvious manner to show that every topological space induced by a relaxed$_\mathrm{t}$ metric is regular and has a basis that can be decomposed into an at most countable collection of neighborhood-finite families. It follows that a topological space is metrizable if and only if it is induced by a c-relaxed$_\mathrm{t}$ metric. That is, every c-relaxed$_\mathrm{t}$ metric is a topological metric.

We now show that the converse fails even if we restrict attention to distance measures (binary nonnegative functions d that are symmetric and satisfy $d(x, y) = 0$ if and only if $x = y$). Define d on the space $[1, \infty)$ by taking $d(x, y) = |y - x|^{\max\{x,y\}}$. It is not hard to verify that d induces the same topology as the usual metric d' with $d'(x, y) = |x - y|$. The intuition is that (1) the ϵ-ball $\{y \mid d(x, y) < \epsilon\}$ is just a minor distortion of an ϵ-ball $\{y \mid d_m(x, y) < \epsilon\}$, where $d_m(x, y) = |x - y|^m$ for some m that depends on x (in fact, with $m = x$), and (2) the function d_m locally induces the same topology as the usual metric d' with $d'(x, y) = |x - y|$. Condition (2) holds since the ball $\{y \mid |x - y|^m < \epsilon\}$ is the same as the ball $\left\{y \mid |x - y| < \epsilon^{1/m}\right\}$. So d is a topological metric. We now show that d is not a c-relaxed$_\mathrm{t}$ metric.

Let $x = 1$, $y = n + 1$, and $z = 2n + 1$. We shall show that for each constant c, there is n such that

$$d(x, z) > c(d(x, y) + d(y, z)). \tag{18}$$

This implies that d is not a relaxed$_\mathrm{t}$ metric. When we substitute for x, y, z in (18), we obtain

$$(2n + 1)^{2n+1} > c((n + 1)^{n+1} + (n + 1)^{2n+1}). \tag{19}$$

But it is easy to see that (19) holds for every sufficiently large n. □

Thus, we have METRIC ⇒ c-RELAXED$_\mathrm{p}$ METRIC ⇒ c-RELAXED$_\mathrm{t}$ METRIC ⇒ TOPOLOGICAL METRIC, and none of the reverse implications hold.

5. Relationships between measures. We now come to the second main result of the paper, where we show that all of our distance measures we have discussed are in the same equivalence class, that is, are bounded by constant multiples of each other both above and below. The connections are proved via two proof methods. We use direct counting arguments to relate F^* with $F_{\min}$, to relate the $K^{(p)}$ measures with each other, and to relate the $F^{(\ell)}$ measures with each other. The more subtle connection between $K_{\min}$ and $F_{\min}$—which provides the link between the measures based on Kendall's tau and the measures based on Spearman's footrule—is proved by applying Diaconis and Graham's inequalities (17) for permutations σ_1, σ_2.

THEOREM 5.1 (main result 2). *The distance measures $K_{\min}$, K_{avg}, K_{Haus}, $K^{(p)}$ (for every choice of p), $F_{\min}$, F_{avg}, F_{Haus}, and $F^{(\ell)}$ (for every choice of ℓ) are all in the same equivalence class.*

The fact that $F^{(\ell)}$ is a metric now implies that all our distance measures are near metrics.

COROLLARY 5.2. *Each of $K^{(p)}$ and $F_{\min}$ (thus also $K_{\min}, K_{\mathrm{avg}}, K_{\mathrm{Haus}}, F_{\mathrm{avg}}, F_{\mathrm{Haus}}$) is a near metric.*

We discuss the proof of this theorem shortly. We refer to the equivalence class that contains all of these distance measures as the *big equivalence class*. The big equivalence class seems to be quite robust. As we have seen, some members of the big equivalence class are metrics.

In later sections, we shall find it convenient to deal with normalized versions of our distance measures by dividing each distance measure by its maximum value. The normalized version is then a distance measure that lies in the interval $[0,1]$.[4] The normalized version is a metric if the original version is a metric, and is a near metric if the original version is a near metric. It is easy to see that if two distance measures are in the same equivalence class, then so are their normalized versions.

Theorem 5.1 is proven by making use of the following theorem (Theorem 5.3), along with Propositions 3.3, 3.4, and 3.7. The bounds in Theorem 5.3 are not tight; while we have improved some of them with more complicated proofs, our goal here is simply to prove enough to obtain Theorem 5.1. If we really wished to obtain tight results, we would have to compare every pair of the distance measures we have introduced, such as $K^{(p)}$ versus $F^{(\ell)}$ for arbitrary p, ℓ.

THEOREM 5.3. *Let τ_1, τ_2 be top k lists.*

(1) $K_{\min}(\tau_1,\tau_2) \le F_{\min}(\tau_1,\tau_2) \le 2K_{\min}(\tau_1,\tau_2)$;

(2) $F^*(\tau_1,\tau_2) \le F_{\min}(\tau_1,\tau_2) \le 2F^*(\tau_1,\tau_2)$;

(3) $K^{(p)}(\tau_1,\tau_2) \le K^{(p')}(\tau_1,\tau_2) \le (\frac{1+p'}{1+p})K^{(p)}(\tau_1,\tau_2)$ *for* $0 \le p \le p' \le 1$;

(4) $F^{(\ell)}(\tau_1,\tau_2) \le F^{(\ell')}(\tau_1,\tau_2) \le (\frac{\ell'-k}{\ell-k})F^{(\ell)}(\tau_1,\tau_2)$ *for* $k < \ell \le \ell'$.

Proof. (1) For the first inequality of part (1), let σ_1, σ_2 be permutations so that $\sigma_1 \succeq \tau_1$, $\sigma_2 \succeq \tau_2$, and $F_{\min}(\tau_1,\tau_2) = F(\sigma_1,\sigma_2)$. Then $F_{\min}(\tau_1,\tau_2) = F(\sigma_1,\sigma_2) \ge K(\sigma_1,\sigma_2) \ge K_{\min}(\tau_1,\tau_2)$, using the first inequality in (17) and the fact that $K_{\min}$ is the minimum over all extensions σ_1 of τ_1 and σ_2 of τ_2.

For the second inequality of part (1), let σ_1, σ_2 be permutations so that $\sigma_1 \succeq \tau_1$, $\sigma_2 \succeq \tau_2$, and $K_{\min}(\tau_1,\tau_2) = K(\sigma_1,\sigma_2)$. Then $K_{\min}(\tau_1,\tau_2) = K(\sigma_1,\sigma_2) \ge (1/2)F(\sigma_1,\sigma_2) \ge (1/2)F_{\min}(\tau_1,\tau_2)$ using the second inequality in (17) and the fact that $F_{\min}$ is minimum over all extensions σ_1 of τ_1 and σ_2 of τ_2.

(2) Let σ_1, σ_2 be permutations so that $\sigma_1 \succeq \tau_1$, $\sigma_2 \succeq \tau_2$, and $F_{\min}(\tau_1,\tau_2) = F(\sigma_1,\sigma_2)$. For $s \in \{1,2\}$, let v_s be a vector such that $v_s(i) = \tau_s(i)$ if $i \in D_{\tau_s}$ and $v_s(i) = k+1$ otherwise. Given τ_1, τ_2, recall that $F^*(\tau_1,\tau_2)$ is exactly the L_1 distance between the corresponding vectors v_1, v_2. If $i \in Z = D_{\tau_1} \cap D_{\tau_2}$, then $|v_1(i) - v_2(i)| = |\sigma_1(i) - \sigma_2(i)|$. If $i \in S = D_{\tau_1} \setminus D_{\tau_2}$, then $|v_1(i) - v_2(i)| = |\tau_1(i) - (k+1)| = |\sigma_1(i) - (k+1)| \le |\sigma_1(i) - \sigma_2(i)|$, since $\sigma_2(i) \ge k+1 > \tau_1(i) = \sigma_1(i)$. The case of $i \in T = D_{\tau_2} \setminus D_{\tau_1}$ is similar. Thus, for every i, we have $|v_1(i) - v_2(i)| \le |\sigma_1(i) - \sigma_2(i)|$. It follows by definition that $F^*(\tau_1,\tau_2) \le F(\sigma_1,\sigma_2) = F_{\min}(\tau_1,\tau_2)$. This proves the first inequality.

We now prove the second inequality. First, we have

$$F_{\min}(\tau_1,\tau_2) = \sum_{i \in Z} |\sigma_1(i) - \sigma_2(i)| + \sum_{i \in S} |\sigma_1(i) - \sigma_2(i)| + \sum_{i \in T} |\sigma_1(i) - \sigma_2(i)|. \tag{20}$$

On the other hand, we have

$$F^*(\tau_1,\tau_2) = \sum_{i \in Z} |\tau_1(i) - \tau_2(i)| + \sum_{i \in S} |\tau_1(i) - (k+1)| + \sum_{i \in T} |(k+1) - \tau_2(i)|. \tag{21}$$

Furthermore, if $z = |Z|$, note that

[4] For metrics on permutations, such as Kendall's tau and Spearman's footrule, it is standard to normalize them to lie in the interval $[-1,1]$, with -1 corresponding to the situation where the permutations are the reverse of each other and with 1 corresponding to the situation where the permutations are equal. However, this normalization immediately precludes one from studying metric-like properties.

$$
\begin{aligned}
\sum_{i\in S}|\tau_1(i)-(k+1)| &\geq \sum_{r=z+1}^{k}|r-(k+1)| \\
&= (k-z)+\cdots+1 \\
&= \frac{(k-z)(k-z+1)}{2}. \qquad (22)
\end{aligned}
$$

By symmetry, we also have $\sum_{i\in T}|(k+1)-\tau_2(i)| \geq (k-z)(k-z+1)/2$.

For $i \in Z$, we have $|\sigma_1(i)-\sigma_2(i)| = |\tau_1(i)-\tau_2(i)|$, and so

$$\sum_{i\in Z}|\sigma_1(i)-\sigma_2(i)| = \sum_{i\in Z}|\tau_1(i)-\tau_2(i)|. \qquad (23)$$

Since $\sigma_2(i) \geq k+1$ and $\tau_1(i) \leq k$ if and only if $i \in S$, we have, for $i \in S$, that $|\tau_1(i)-\sigma_2(i)| = |\tau_1(i)-(k+1)| + (\sigma_2(i)-(k+1))$. Furthermore, since σ_2 is a permutation, the list of values $\sigma_2(i), i \in S$, is precisely $k+1,\ldots,2k-z$. Summing over all $i \in S$ yields

$$
\begin{aligned}
\sum_{i\in S}|\sigma_1(i)-\sigma_2(i)| &= \sum_{i\in S}|\tau_1(i)-\sigma_2(i)| \\
&= 0+1+\cdots+(k-z-1)+\sum_{i\in S}|\tau_1(i)-(k+1)| \\
&= \frac{(k-z-1)(k-z)}{2}+\sum_{i\in S}|\tau_1(i)-(k+1)| \\
&\leq 2\sum_{i\in S}|\tau_1(i)-(k+1)| \qquad \text{by (22).} \qquad (24)
\end{aligned}
$$

Similarly, we also have

$$\sum_{i\in T}|\sigma_1(i)-\sigma_2(i)| \leq 2\sum_{i\in T}|(k+1)-\tau_2(i)|. \qquad (25)$$

Now, using (20)–(25), we have $F_{\min}(\tau_1,\tau_2) \leq 2F^*(\tau_1,\tau_2)$.

(3) From the formula given in Lemma 3.1, we have

$$K^{(p')}(\tau_1,\tau_2) - K^{(p)}(\tau_1,\tau_2) = (k-z)(p'-p)(k-z-1). \qquad (26)$$

The first inequality is immediate from (26), since $k \geq z$.

We now prove the second inequality. If $K^{(p)}(\tau_1,\tau_2) = 0$, then $\tau_1 = \tau_2$, so also $K^{(p')}(\tau_1,\tau_2) = 0$, and the second inequality holds. Therefore, assume that $K^{(p)}(\tau_1,\tau_2) \neq 0$. Divide both sides of (26) by $K^{(p)}(\tau_1,\tau_2)$ to obtain

$$\frac{K^{(p')}(\tau_1,\tau_2)}{K^{(p)}(\tau_1,\tau_2)} = 1 + \frac{(k-z)(p'-p)(k-z-1)}{K^{(p)}(\tau_1,\tau_2)}. \qquad (27)$$

Since $\frac{1+p'}{1+p} = 1 + \frac{p'-p}{1+p}$, the second inequality would follow from (27) if we show

$$K^{(p)}(\tau_1,\tau_2) \geq (k-z)(k-z-1)(1+p). \qquad (28)$$

In the derivation of the formula for $K^{(p)}(\tau_1,\tau_2)$ in the proof of Lemma 3.1, we saw that the contribution from Case 3 is $(k-z)^2$ and the contribution from Case 4 is $p(k-z)(k-z-1)$. Hence, $K^{(p)}(\tau_1,\tau_2) \geq (k-z)^2 + p(k-z)(k-z-1) \geq (k-z)(k-z-1) + p(k-z)(k-z-1) = (k-z)(k-z-1)(1+p)$, as desired.

(4) From the formula given in Lemma 3.5, we have

$$F^{(\ell')}(\tau_1,\tau_2) - F^{(\ell)}(\tau_1,\tau_2) = 2(k-z)(\ell'-\ell). \tag{29}$$

The first inequality is immediate from (29), since $k \geq z$.

We now prove the second inequality. If $F^{(\ell)}(\tau_1,\tau_2) = 0$, then $\tau_1 = \tau_2$, so also $F^{(\ell')}(\tau_1,\tau_2) = 0$, and the second inequality holds. Therefore, assume that $F^{(\ell)}(\tau_1,\tau_2) \neq 0$. Divide both sides of (29) by $F^{(\ell)}(\tau_1,\tau_2)$ to obtain

$$\frac{F^{(\ell')}(\tau_1,\tau_2)}{F^{(\ell)}(\tau_1,\tau_2)} = 1 + \frac{2(k-z)(\ell'-\ell)}{F^{(\ell)}(\tau_1,\tau_2)}. \tag{30}$$

Since $\frac{\ell'-k}{\ell-k} = 1 + \frac{\ell'-\ell}{\ell-k}$, the second inequality would follow from (30) if we show

$$F^{(\ell)}(\tau_1,\tau_2) \geq 2(k-z)(\ell-k). \tag{31}$$

To see (31), observe that $|S|+|T| = 2(k-z)$ and each element in S and T contributes at least $\ell-k$ (which is positive since $k<\ell$) to $F^{(\ell)}(\tau_1,\tau_2)$. □

6. An algorithmic application. In the context of algorithm design, the notion of near metrics has the following useful application. Given r ranked lists $\tau_1,\ldots,\tau_r$ (either full lists or top k lists) of "candidates," the *rank aggregation* problem [DKNS01] with respect to a distance measure d is to compute a list τ (again, either a full list or another top k list) such that $\sum_{j=1}^{r} d(\tau_j,\tau)$ is minimized.

This problem arises in the context of IR, where possible results to a search query may be ordered with respect to several criteria, and it is useful to obtain an ordering (often a top k list) that is a good aggregation of the rank orders produced. It is argued in [DKNS01] that Kendall's tau and its variants are good measures to use, both in the context of full lists and top k lists. Our experiments at the IBM Almaden Research Center (see also section 9.1) have confirmed that, in fact, producing an ordering with small Kendall's tau distance yields qualitatively excellent results. Unfortunately, computing an optimal aggregation of several full or top k lists is NP-hard for each of the Kendall measures. In this context, our notion of an equivalence class of distance measures comes in handy.

PROPOSITION 6.1. *Let $\mathcal{C}$ be an equivalence class of distance measures. If there is at least one distance measure d in $\mathcal{C}$ so that the rank aggregation problem with respect to d has a polynomial-time exact or constant-factor approximation algorithm, then for every d' in $\mathcal{C}$, there is a polynomial-time constant-factor approximation algorithm for the rank aggregation problem with respect to d'.*

Proof. Given $\tau_1,\ldots,\tau_r$, let τ denote an aggregation with respect to d that is within a factor $c \geq 1$ of a best possible aggregation π with respect to d, that is, $\sum_j d(\tau_j,\tau) \leq c\sum_j d(\tau_j,\pi)$. Let c_1, c_2 denote positive constants such that for all σ,σ' (top k or full lists, as appropriate) $c_1 d(\sigma,\sigma') \leq d'(\sigma,\sigma') \leq c_2 d(\sigma,\sigma')$. Also, let π' denote a best possible aggregation with respect to d'. Then we have

$$\begin{aligned}\sum_j d'(\tau_j,\tau) &\leq \sum_j c_2 d(\tau_j,\tau) \leq c\sum_j c_2 d(\tau_j,\pi)\\ &\leq cc_2\sum_j d(\tau_j,\pi') \leq \frac{cc_2}{c_1}\sum_j d'(\tau_j,\pi'). \qquad \square\end{aligned}$$

Via an application of minimum-cost perfect matching, the rank aggregation problem can be solved optimally in polynomial time for any of the $F^{(\ell)}$ metrics. Together with Theorem 5.1, this implies polynomial-time constant-factor approximation algorithms for the rank aggregation problem with respect to the Kendall measures.

7. Other approaches.

7.1. Spearman's rho. Spearman's rho is the L_2 distance between two permutations. Formally,

$$\rho(\sigma_1, \sigma_2) = \left(\sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)|^2 \right)^{1/2}$$

and it can be shown that $\rho(\cdot, \cdot)$ is a metric.[5] The maximum value of $\rho(\sigma_1, \sigma_2)$ is $(n(n+1)(2n+1)/3)^{\frac{1}{2}}$, which occurs when σ_1 is the reverse of σ_2. Spearman's rho is a popular metric between permutations. Analogous to the footrule case, we can define the notions of ρ_{min}, ρ_{avg}, and $\rho^{(\ell)}$. They are not in the big equivalence class for the following reason. Consider the case where $k = n$, that is, where we are considering full lists, which are permutations of all of the elements in a fixed universe. In this case, we need only consider ρ, since ρ_{min}, ρ_{avg}, and $\rho^{(\ell)}$ all equal ρ. But the maximum value of F^* is $\Theta(n^2)$ and that of ρ is $\Theta(n^{\frac{3}{2}})$. Therefore, ρ_{min}, ρ_{avg}, and $\rho^{(\ell)}$ cannot be in the same equivalence class as F^*. What if we consider normalized versions of our distance measures, as discussed after Theorem 5.1? We now show that the normalized versions of ρ_{min}, ρ_{avg}, and $\rho^{(\ell)}$ are not in the normalized version of the big equivalence class. If d is a distance measure, we will sometimes denote the normalized version of d by $\dot{d}$.

PROPOSITION 7.1. *The distance measures ρ_{min}, ρ_{avg}, and $\rho^{(\ell)}$ do not belong to the big equivalence class, even if all distance measures are normalized.*

Proof. As before, we consider full lists. We will show that $\dot{F}^*$ and $\dot{\rho}$ do not bound each other by constant multiples. We will present a family of pairs of full lists, one for each n, such that $\dot{F}^*(\sigma_1, \sigma_2) = \Theta(1/n)$ and $\dot{\rho}(\sigma_1, \sigma_2) = \Theta(1/n^{\frac{3}{4}})$. For every n, let $r = \lceil \sqrt{n} \rceil$. Assume n is large enough so that $n \geq 2r$. Define the permutation σ_1 so that the elements in order are $1, \ldots, n$, and define the permutation σ_2 so that the elements in order are $r+1, \ldots, 2r, 1, \ldots, r, 2r+1, \ldots, n$. The unnormalized versions of Spearman's footrule and Spearman's rho can be easily calculated to be $F^*(\sigma_1, \sigma_2) = 2r^2 = \Theta(n)$ and $\rho(\sigma_1, \sigma_2) = (2r)^{\frac{3}{2}} = \Theta(n^{\frac{3}{4}})$. As we noted, the maximum value of F^* is $\Theta(n^2)$ and that of ρ is $\Theta(n^{\frac{3}{2}})$. Therefore, $\dot{F}^*(\sigma_1, \sigma_2) = \Theta(1/n)$ and $\dot{\rho}(\sigma_1, \sigma_2) = \Theta(1/n^{\frac{3}{4}})$. Thus $\dot{F}^*$ and $\dot{\rho}$ cannot bound each other by constant multiples, so $\dot{\rho}_{\text{min}}$, $\dot{\rho}_{\text{avg}}$, and $\dot{\rho}^{(\ell)}$ do not belong to the normalized version of the big equivalence class. □

7.2. The intersection metric. A natural approach to defining the distance between two top k lists τ_1 and τ_2 is to capture the extent of overlap between D_{τ_1} and D_{τ_2}. We now define a more robust version of this distance measure. For $1 \leq i \leq k$, let $\tau^{(i)}$ denote the restriction of a top k list to the first i items. Let

$$\delta_i^{(w)}(\tau_1, \tau_2) = |D_{\tau_1^{(i)}} \Delta D_{\tau_2^{(i)}}|/(2i).$$

Finally, let

[5] Spearman's rho is usually defined without the exponent of $\frac{1}{2}$, that is, without the square root. However, if we drop the exponent of $\frac{1}{2}$, then the resulting distance measure is not a metric, and is not even a near metric.

$$\delta^{(w)}(\tau_1, \tau_2) = \frac{1}{k} \sum_{i=1}^{k} \delta_i^{(w)}(\tau_1, \tau_2).$$

(Here, Δ represents the symmetric difference. Thus, $X \Delta Y = (X \setminus Y) \cup (Y \setminus X)$.) It is straightforward to verify that $\delta^{(w)}$ lies between 0 and 1, with the maximal value of 1 occurring when D_{τ_1} and D_{τ_2} are disjoint. In fact, $\delta^{(w)}$, as defined above, is just one instantiation of a more general paradigm: any convex combination of the $\delta_i^{(w)}$'s with strictly positive coefficients yields a metric on top k lists.

We now show that the distance measure $\delta^{(w)}$ is a metric.

Proposition 7.2. *$\delta^{(w)}(\cdot, \cdot)$ is a metric.*

Proof. It suffices to show that $\delta_i^{(w)}(\cdot, \cdot)$ is a metric for $1 \le i \le k$. To show this, we show that for any three sets A, B, C, we have $|A \Delta C| \le |A \Delta B| + |B \Delta C|$. For $x \in A \Delta C$, assume without loss of generality that $x \in A$ and $x \notin C$. We have two cases: if $x \in B$, then $x \in B \Delta C$ and if $x \notin B$, then $x \in A \Delta B$. Either way, each $x \in A \Delta C$ contributes at least one to the right-hand side, thus establishing the inequality. □

Since $\delta^{(w)}$ is bounded (by 1), and F^* is not bounded, it follows that $\delta^{(w)}$ is not in the big equivalence class. Of course, $\delta^{(w)}$ is normalized; we now show that $\delta^{(w)}$ is not in the normalized version of the big equivalence class.

Proposition 7.3. *$\delta^{(w)}$ does not belong to the equivalence class, even if all distance measures are normalized.*

Proof. Let τ_1 be the top k list where the top k elements in order are $1, 2, \ldots, k$, and let τ_2 be the top k list where the top k elements in order are $2, \ldots, k, 1$. The normalized footrule can be calculated to be $\dot{F}^*(\tau_1, \tau_2) = \Theta(1/k)$, whereas $\delta^{(w)}(\tau_1, \tau_2) = (1/k) \sum_{i=1}^{k} 1/i = \Theta((\ln k)/k)$. Therefore, $\delta^{(w)}$ and $\dot{F}^*$ cannot bound each other by constant multiples, and so $\delta^{(w)}$ does not belong to the normalized version of the big equivalence class. □

7.3. Goodman and Kruskal's gamma. Goodman and Kruskal [GK54] have defined a "correlation statistic" for rank orders (and partial orders), which can be used to define a distance measure for top k lists. Let τ_1 and τ_2 be top k lists. As before, let $\mathcal{P}(\tau_1, \tau_2) = \mathcal{P}_{D_{\tau_1} \cup D_{\tau_2}}$ be the set of all unordered pairs of distinct elements in $D_{\tau_1} \cup D_{\tau_2}$. Let C be the set of all pairs $\{i, j\} \in \mathcal{P}(\tau_1, \tau_2)$ where both τ_1 and τ_2 implicitly or explicitly place one of i or j above the other (τ_1 and τ_2 can differ on this placement). In other words, C consists of all pairs $\{i, j\} \in \mathcal{P}(\tau_1, \tau_2)$ such that (1) either i or j is in D_{τ_1} and (2) either i or j is in D_{τ_2}. Note that C consists exactly of all pairs $\{i, j\}$ that occur in the first three cases in our definition of $K^{(p)}$. Now define $\gamma(\tau_1, \tau_2)$ to be the fraction of pairs $\{i, j\} \in C$ where τ_1 and τ_2 disagree on whether i is ahead of j.

Goodman and Kruskal defined this quantity for rank orders τ_1 and τ_2 that are more general than top k lists, namely, "bucket orders," or total orders with ties.[6] However, this quantity is not well defined for all pairs of bucket orders, since the set C as defined above can be empty in general. In ongoing work, we are exploring the issue of bucket orders in more detail. Here we simply remark that if τ_1 and τ_2 are top k lists, then C is always nonempty, and so we do obtain a meaningful distance measure on top k lists via this approach.

[6] As with Kendall's tau and Spearman's footrule (see footnote 4), Goodman and Kruskal's gamma is traditionally normalized to lie in the interval $[-1, 1]$, although we shall not do so, so that we can discuss metric properties.

We now show that γ is not a metric. Let τ_1 be the top 4 list where the top 4 items in order are 1,2,3,4; let τ_2 be the top 4 list where the top 4 items in order are 1,2,5,6; and let τ_3 be the top 4 list where the top 4 items in order are 5,6,7,8. It is straightforward to verify that $\gamma(\tau_1, \tau_3) = 1$, $\gamma(\tau_1, \tau_2) = 4/13$, and $\gamma(\tau_2, \tau_3) = 8/13$. So the triangle inequality fails, because $\gamma(\tau_1, \tau_3) > \gamma(\tau_1, \tau_2) + \gamma(\tau_2, \tau_3)$.

We now show that γ belongs to the normalized version of our big equivalence class and is therefore a near metric. Let τ_1 and τ_2 be top k lists, and let C be as earlier. Let $c = |C|$, and let s be the number of pairs $\{i, j\} \in C$ where τ_1 and τ_2 disagree on whether i is ahead of j. Thus, $\gamma(\tau_1, \tau_2) = s/c$. Note that since $c \leq k^2$, we have $s/c \geq s/k^2 = K_{\min}(\tau_1, \tau_2)/k^2$, which equals the normalized $K_{\min}$ distance between τ_1 and τ_2. Finally, note that since $c \geq \binom{k}{2}$, we have $s/c \leq s/\binom{k}{2} \leq 4s/k^2$ (for $k \geq 2$). Therefore, s/c is at most 4 times the normalized $K_{\min}$ distance between τ_1 and τ_2 if $k \geq 2$. (It is easy to see that γ and the normalized version of $K_{\min}$ are both 0 or both 1 when $k = 1$.)

8. The interpolation criterion.

In practical situations where one compares two top k lists, it would be nice if the distance value has some natural real-life interpretation associated with it. There are three possible extreme relationships between two top k lists: (a) they are identical, (b) they contain the same k elements in the exact opposite order, or (c) they are disjoint. We feel that it is desirable that the value in case (b) be about halfway between the values in cases (a) and (c).

Let d denote any one of our distance measures between top k lists τ_1 and τ_2. Analogous to the normalization given in footnote 4 of section 5, let us obtain a normalized version ν that maps the distance values into the interval $[-1, 1]$ so that

(a) $\nu(\tau_1, \tau_2) = 1$ if and only if $\tau_1 = \tau_2$;

(b) $\nu(\tau_1, \tau_2) = -1$ if and only if D_{τ_1} and D_{τ_2} are disjoint, that is, $Z = \emptyset$.

Clearly, this can be achieved via a linear map of the form $\nu(\tau_1, \tau_2) = a \cdot d(\tau_1, \tau_2) + b$. The question now is, How close to zero is $\nu(\tau_1, \tau_2)$ when τ_1 and τ_2 contain the same k elements in the exact opposite order?

It turns out that the answer is asymptotic (in k) to $p/(1+p)$ for $K^{(p)}$. Therefore, it is asymptotic to 0 for $K_{\min} = K^{(0)}$. In fact, for $K_{\min}$, it is $\Theta(1/k)$. For $F_{\min}$, it is $\frac{1}{2}$, and for $F^{(\ell)}$, with $\ell = k + \frac{1}{2} + \alpha$, it is $\Theta(\frac{\alpha}{k+\alpha})$. In fact, for $F^{(k+\frac{1}{2})}$, where $\alpha = 0$, it is $\Theta(1/k^2)$. Thus, from this viewpoint, the preferable distance measures are $K_{\min}$ and $F^{(k+\beta)}$ for $\beta = o(k)$ (which includes F^*).

9. Experiments.

9.1. Comparing Web search engines.

As we mentioned earlier, one of the important applications of comparing top k lists is to provide an objective way to compare the output of different search engines. We illustrate the use of our methods by comparing the outputs of seven popular Web search engines: AltaVista (www.altavista.com), Lycos (www.lycos.com), AllTheWeb (www.alltheweb.com), HotBot (www.hotbot.com), NorthernLight (www.northernlight.com), AOL Search (search.aol.com), and MSN Search (search.msn.com). Comparing the output in this manner will shed light both on the similarities between the underlying indices and the ranking functions used by search engines. We selected $K_{\min}$ as the measure of comparison between the search engines. This choice is arbitrary, and as we argued earlier, we could just as well have chosen any other measure from the big equivalence class.

We made use of 750 queries, that were actually made by real users to a metasearch engine developed at the IBM Almaden Research Center [DKNS01]. For each of these queries, and for each of the seven Web search engines we are considering, we obtained

Table 1
$K_{\min}$ distances between search engines for $k = 50$.

	AltaVista	Lycos	AllTheWeb	HotBot	NorthernLight	AOL Search	MSN Search
AltaVista	0.000	0.877	0.879	0.938	0.934	0.864	0.864
Lycos	0.877	0.000	0.309	0.888	0.863	0.796	0.790
AllTheWeb	0.879	0.309	0.000	0.873	0.866	0.782	0.783
HotBot	0.938	0.888	0.873	0.000	0.921	0.516	0.569
NorthernLight	0.934	0.863	0.866	0.921	0.000	0.882	0.882
AOL Search	0.864	0.796	0.782	0.516	0.882	0.000	0.279
MSN Search	0.864	0.790	0.783	0.569	0.882	0.279	0.000

the top 50 list.[7] We then computed the normalized $K_{\min}$ distance between every pair of search engine outputs. Finally, we averaged the distances over the 750 queries. The results are tabulated in Table 1. The values are normalized to lie between 0 and 1, with smaller values representing closer matches. Note, of course, that the table is symmetric about the main diagonal.

Several interesting conclusions can be derived from this table. Some of the conclusions are substantiated by the alliances between various search engines. (For a detailed account of the alliances, see www.searchenginewatch.com/reports/alliances.html.)

(1) AOL Search and MSN Search yield very similar results! The reason for this (surprising) behavior is twofold: both AOL Search and MSN Search index similar sets of pages and probably use fairly similar ranking functions. These conclusions are substantiated by the fact that AOL Search uses search data from OpenDirectory and Inktomi, and MSN Search uses LookSmart and Inktomi. HotBot uses DirectHit and Inktomi and can be seen to be moderately similar to AOL Search and MSN Search.

(2) Lycos and AllTheWeb yield similar results. Again, the reason for this is because Lycos gets its main results from DirectHit and AllTheWeb.

(3) AltaVista and NorthernLight, since they use their own crawling, indexing, and ranking algorithms, are far away from every other search engine. This is plausible for two reasons: either they crawl and index very different portions of the Web or their ranking functions are completely unrelated to the ranking functions of the other search engines.

(4) The fact that $K_{\min}$ is a near metric allows us to draw additional interesting inferences from the tables (together with observations (1) and (2) above). For example, working through the alliances and partnerships mentioned above, and exploiting the transitivity of "closeness" for a near metric, we obtain the following inference. The data services LookSmart and OpenDirectory are closer to each other than they are to DirectHit. Given that DirectHit uses results from its own database and from OpenDirectory, this suggests that the in-house databases in DirectHit and OpenDirectory are quite different. A similar conclusion is again supported by the fact that Lycos and HotBot are far apart, and their main results are powered by OpenDirectory and DirectHit, respectively.

9.2. Evaluating a metasearch engine. Recall that a metasearch engine combines the ranking of different search engines to produce an aggregated ranking. There are several metasearch engines available on the Web (for a list of popular ones, see the site searchenginewatch.com). Metasearch engines are quite popular for their coverage, resistance to spam, and ability to mitigate the quirks of crawl. As we mentioned earlier, our methods can be used to evaluate the behavior of a metasearch engine. Such

[7] For some queries, we had to work with a slightly smaller value of k than 50, since a search engine returned some duplicates.

TABLE 2
$K_{\min}$ distance of our metasearch engine to its sources for $k = 50$.

AltaVista	Lycos	AllTheWeb	HotBot	NorthernLight	AOL Search	MSN Search
0.730	0.587	0.565	0.582	0.823	0.332	0.357

an analysis will provide evidence to whether the metasearch is highly biased towards any particular search engine or is reasonably "close" to all the search engines.

For our purposes, we use a metasearch engine that we developed. Our metasearch engine uses a Markov chain approach to aggregate various rankings. The underlying theory behind this method can be found in [DKNS01]. We used a version of our metasearch engine that combines the outputs of the seven search engines described above. We measured the average $K_{\min}$ distance of our metasearch engine's output to the output of each of the search engines for the same set of 750 queries. The results are tabulated in Table 2. From this table and Table 1, we note the following. There is a strong bias towards the AOL Search/MSN Search cluster, somewhat less bias towards Lycos, AllTheWeb, and HotBot, and very little bias towards AltaVista and NorthernLight. This kind of information is extremely valuable for metasearch design (and is beyond the scope of this paper). For example, the numbers show that the output of our metasearch engine is a reasonable aggregation of its sources—it does not simply copy its components, nor does it exclude any component entirely. Finally, the degree to which our metasearch engine aligns itself with a search engine depends on the various reinforcements among the outputs of the search engines.

9.3. Correlations among the distance measures. The following experiment is aimed at studying the "correlations" between the distance measures. We seek to understand how much information the distance measures reveal about each other. One of the goals of this experiment is to find empirical support for the following belief motivated by our work in this paper: the distance measures within an equivalence class all behave similarly, whereas different equivalence classes aim to capture different aspects of the distance between two lists.

Let I denote the top k list where the top k elements in order are $1, 2, \ldots, k$. For a distance measure $d(\cdot, \cdot)$ and a top k list τ with elements from the universe $\{1, 2, \ldots, 2k\}$, let $\hat{d}(\tau) = d(\tau, I)$. If τ is a randomly chosen top k list, then $\hat{d}(\tau)$ is a random variable.

Let d_1 and d_2 denote two distance measures. Consider the experiment where a random top k list τ is picked. Informally, the main question we ask here is the following: if we know $\widehat{d_1}(\tau)$ (namely, the distance, according to d_1, of τ to the list I), to what extent can we predict the value of $\widehat{d_2}(\tau)$? To address this question, we use two basic notions from information theory.

Recall that the entropy of a random variable X is

$$H(X) = -\sum_x \Pr[X = x] \log \Pr[X = x].$$

If we truncate the precision to two digits and use logarithms to the base 10 in the entropy definition, then for each d, the quantity $H(\hat{d}(\tau))$ is a real number between 0 and 2. In words, when τ is picked at random, then there is up to "2 digits worth of uncertainty in the value of $\hat{d}(\tau)$."

Table 3

Conditional entropy values for pairs of distance measures. The entry (d_1, d_2) of the table may be interpreted as the average uncertainty in $\widehat{d_2}(\tau)$, assuming we know $\widehat{d_1}(\tau)$.

	δ	$\delta^{(w)}$	$\rho^{(k+1)}$	γ	F^*	$F_{\min}$	$K_{\min}$	K_{avg}	$K^{(1)}$
δ	0.000	1.409	1.469	1.415	1.203	1.029	1.235	1.131	0.991
$\delta^{(w)}$	0.580	0.000	1.193	1.282	0.863	0.945	1.087	1.091	1.043
$\rho^{(k+1)}$	0.530	1.083	0.000	1.057	0.756	0.834	0.670	0.773	0.760
γ	0.503	1.197	1.082	0.000	1.039	1.025	0.533	0.525	0.507
F^*	0.497	0.985	0.989	1.246	0.000	0.434	0.848	0.845	0.819
$F_{\min}$	0.388	1.132	1.131	1.297	0.499	0.000	0.885	0.748	0.650
$K_{\min}$	0.490	1.170	0.863	0.700	0.808	0.780	0.000	0.454	0.500
K_{avg}	0.421	1.210	1.002	0.729	0.841	0.680	0.490	0.000	0.354
$K^{(1)}$	0.361	1.240	1.068	0.789	0.894	0.660	0.615	0.433	0.000

The conditional entropy of a random variable X with respect to another random variable Y is

$$H(X \mid Y) = \sum_y \Pr[Y = y] H(X \mid Y = y).$$

Informally, the conditional entropy measures the uncertainty in X, assuming that we know the value of Y. In our case, we ask the question: For a random τ, if we know the value of $\widehat{d_1}(\tau)$, how much uncertainty is left in the value of $\widehat{d_2}(\tau)$?[8]

For all pairs of our distance measures d_1 and d_2, we measure $H(\widehat{d_2}(\tau) \mid \widehat{d_1}(\tau))$, and present the results in Table 3. We consider a universe of 20 elements and let $k = 10$. (These choices enable us to exhaustively enumerate all possible top k lists and perform our experiments on them.) The entry (d_1, d_2) in this table denotes $H(\widehat{d_2}(\tau) \mid \widehat{d_1}(\tau))$. Therefore, the closer the value is to 2, the less information $\widehat{d_1}$ reveals about $\widehat{d_2}$. The value of 1 is an interesting case, since this roughly corresponds to saying that on the average, given $\widehat{d_1}(\tau)$, one can predict the leading digit of $\widehat{d_2}(\tau)$.

Some conclusions that can be drawn from the table are the following:

(1) Every distance measure reveals a lot of information about symmetric difference δ. A reason for this is that δ uses only 10 distinct values between 0 and 1, and is not sharp enough to yield finer information. This suggests that the other measures are preferable to symmetric difference.

(2) The distance measure $\rho^{(k+1)}$ reveals much information about the other measures, as is evident from the row for $\rho^{(k+1)}$; on the other hand, as can be seen from the column for $\rho^{(k+1)}$, the other measures do not reveal much information about $\rho^{(k+1)}$. The weighted symmetric difference metric $\delta^{(w)}$ seems fairly unrelated to all the others.

(3) The measures in the big equivalence class all appear to have a stronger correlation to themselves than to the ones not in the class. In fact, each of the footrule measures $F_{\min}, F^*$ is strongly correlated with the other footrule measures, as is evident from the entries corresponding to their submatrix. Similarly, the Kendall measures $K_{\min}, K_{\text{avg}}, K^{(1)}$ are all strongly correlated. This suggests that the footrule and

[8] We chose conditional entropy instead of statistical notions like correlation for the following reason. Correlation (covariance divided by the product of standard deviations) measures linear relationships between random variables. For example, if $X = \alpha Y + \beta$ for some constants α and β, then the correlation between X and Y is zero. On the other hand, consider $X = \alpha Y^2 + \beta Y + \gamma$; even though given the value of Y, there is absolutely no uncertainty in the value of X, their correlation is not zero. Conditional entropy, however, can measure arbitrary functional relationships between random variables. If $X = f(Y)$ for any fixed function f, then $H(X \mid Y) = 0$.

Kendall measures form two "mini"-equivalence classes that sit inside the big equivalence class.

(4) The distance measure γ reveals much information about the Kendall measures, and vice versa. This is to be expected, since γ is very similar to $K_{\min}$, except for the normalization factor.

10. Conclusions. We have introduced various distance measures between top k lists and have shown that these distance measures are equivalent in a very natural sense. We have also introduced a robust notion of "near metric," which we think is interesting in its own right. We have shown that each of our distance measures that is not a metric is a near metric. Our results have implications for IR (since we can quantify the differences between search engines, by measuring the difference between their outputs). Our results also have implications for algorithm design (since we can use our machinery to obtain polynomial-time constant-factor approximation algorithms for the rank aggregation problem).

Acknowledgments. We thank Moni Naor and Gagan Aggarwal for helpful suggestions.

REFERENCES

[AB95] T. Andreae and H.-J. Bandelt, *Performance guarantees for approximation algorithms depending on parametrized triangle inequalities*, SIAM J. Discrete Math., 8 (1995), pp. 1–16.

[BC00] M. A. Bender and C. Chekuri, *Performance guarantees for the TSP with a parameterized triangle inequality*, Inform. Process. Lett., 73 (2000), pp. 17–21.

[CCF$^+$01] D. Carmel, D. Cohen, R. Fagin, E. Farchi, M. Herscovici, Y. Maarek, and A. Soffer, *Static index pruning for information retrieval systems*, in Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval, 2001, pp. 43–50.

[CCF02] M. Charikar, K. Chen, and M. Farach-Colton, *Finding frequent items in data streams*, in Proceedings of the 29th International Colloquium on Automata, Languages, and Programming, Lecture Notes in Comput. Sci. 2380, Springer-Verlag, Berlin, 2002, pp. 693–703.

[Cri80] D. E. Critchlow, *Metric Methods for Analyzing Partially Ranked Data*, Lecture Notes in Statist. 34, Springer-Verlag, Berlin, 1980.

[DG77] P. Diaconis and R. Graham, *Spearman's footrule as a measure of disarray*, J. Roy. Statist. Soc., Ser. B, 39 (1977), pp. 262–268.

[Dia88] P. Diaconis, *Group Representation in Probability and Statistics*, IMS Lecture Notes Monogr. Ser. 11, Institute of Mathematical Statistics, Hayward, CA, 1988.

[DKNS01] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, *Rank aggregation methods for the web*, in Proceedings of the 10th International World Wide Web Conference, ACM, New York, 2001, pp. 613–622.

[Dug66] J. Dugundji, *Topology*, Allyn and Bacon, Boston, 1966.

[FKS03] R. Fagin, R. Kumar, and D. Sivakumar, *Comparing top k lists*, in Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 2003, pp. 28–36.

[FS98] R. Fagin and L. Stockmeyer, *Relaxing the triangle inequality in pattern matching*, Int. J. Comput. Vision, 30 (1998), pp. 219–231.

[GK54] L. A. Goodman and W. H. Kruskal, *Measures of association for cross classifications*, J. Amer. Statist. Assoc., 49 (1954), pp. 732–764.

[KG90] M. Kendall and J. D. Gibbons, *Rank Correlation Methods*, Edward Arnold, London, 1990.

[KHMG03] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub, *Extrapolation methods for accelerating PageRank computations*, in Proceedings of the 12th International World Wide Web Conference, ACM, New York, 2003, pp. 261–270.

[Lee95] J. H. Lee, *Combining multiple evidence from different properties of weighting schemes*, in Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval, 1995, pp. 180–188.

[Lee97] J. H. Lee, *Combining multiple evidence from different relevant feedback methods*, in Database Systems for Advanced Applications '97, World Scientific, Singapore, 1997, pp. 421–430.

SIAM J. DISCRETE MATH.
Vol. 17, No. 2, pp. vii–vii

CORRECTION TO "COMPARING TOP k LISTS"

The beginning of section 3 of "Comparing Top k Lists," *SIAM Journal on Discrete Mathematics*, 17 (2003), pp. 134–160, by Ronald Fagin, Ravi Kumar, and D. Sivakumar, should read as follows:
We now discuss modifications of these metrics for the case when we have only the top k members of the ordering. Formally, a *top k list* τ is a bijection from a domain D_τ (intuitively, the members of the top k list) to $[k]$.

SIAM J. DISCRETE MATH.
Vol. 17, No. 1, pp. 161–169

ACYCLIC HOMOMORPHISMS AND CIRCULAR COLORINGS OF DIGRAPHS[*]

TOMÁS FEDER[†], PAVOL HELL[‡], AND BOJAN MOHAR[§]

Abstract. An acyclic homomorphism of a digraph D into a digraph F is a mapping $\phi\colon V(D) \to V(F)$ such that for every arc $uv \in E(D)$, either $\phi(u) = \phi(v)$ or $\phi(u)\phi(v)$ is an arc of F, and for every vertex $v \in V(F)$, the subgraph of D induced on $\phi^{-1}(v)$ is acyclic. For each fixed digraph F we consider the following decision problem: Does a given input digraph D admit an acyclic homomorphism to F? We prove that this problem is NP-complete unless F is acyclic, in which case it is polynomial time solvable. From this we conclude that it is NP-complete to decide if the circular chromatic number of a given digraph is at most q, for any rational number $q > 1$. We discuss the complexity of the problems restricted to planar graphs. We also refine the proof to deduce that certain F-coloring problems are NP-complete.

Key words. digraph, graph homomorphism, acyclic homomorphism, circular coloring, circular chromatic number, NP-completeness

AMS subject classifications. 05C15, 05C20, 68Q17

DOI. 10.1137/S0895480103422184

1. Introduction.

Let H be a fixed graph. An *H-coloring* of a graph G is a graph homomorphism $G \to H$, i.e., a mapping $\phi\colon V(G) \to V(H)$ such that $\phi(u)\phi(v)$ is an edge of H whenever uv is an edge of G. This notion generalizes k-coloring since a K_k-coloring of G is precisely a standard k-coloring of G. For a fixed integer $k \geq 3$, to decide the existence of a k-coloring for a given graph G is one of the basic NP-complete problems. This result has been generalized to H-colorings by Hell and Nešetřil [10], who proved that, for a fixed graph H, to decide the existence of an H-coloring for a given graph G is NP-complete if H is not bipartite (and is polynomially solvable if H is bipartite).

Let $C(k, d)$ be the graph with vertex set $\{0, \dots, k-1\}$ in which distinct vertices i, j are adjacent if and only if $d \leq |i-j| \leq k-d$. The *circular chromatic number* $\chi_c(G)$ of a graph G is the minimum of all rational numbers k/d (where k and $d \leq k$ are positive integers) such that there exists a homomorphism $G \to C(k, d)$ (the minimum must exist [5]). Thus $\chi_c(G) \leq \frac{k}{d}$ if and only if there exists a homomorphism $G \to C(k, d)$, and the result of [10] implies that for every given rational number $q > 2$, it is also NP-complete to decide if a given graph G has $\chi_c(G) \leq q$. (See also [5, 8, 9]; in particular, it is known to be hard to decide if a graph G with $\chi(G) = n$ has $\chi_c(G) \leq n - \frac{1}{k}$ for any integers $k \geq 2, n \geq 3$ [9].)

The theory of circular colorings of graphs has become an important branch of chromatic graph theory with many exciting results and new techniques. We refer to the survey article by Zhu [15]. Recently, one of the authors [14] has extended the

[*] Received by the editors February 5, 2003; accepted for publication (in revised form) June 30, 2003; published electronically October 14, 2003.
http://www.siam.org/journals/sidma/17-1/42218.html

[†] 268 Waverley St., Palo Alto, CA 94301 (tomas@theory.stanford.edu).

[‡] School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6 (pavol@cs.sfu.ca).

[§] Department of Mathematics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia (bojan.mohar@uni-lj.si).

The work of this author was supported in part by the Ministry of Science and Technology of Slovenia, research project J1-0502-0101-01.

notion of circular colorings to graphs with weighted edges, which can be specialized to also yield a notion of the circular chromatic number $\chi_c(D)$ of a digraph [4].

Let D be a digraph. (All digraphs will be assumed to have no loops.) A vertex set $A \subseteq V(D)$ is *acyclic* if the induced subgraph $D[A]$ is acyclic. A partition of $V(D)$ into k acyclic sets is called a *k-coloring* of D. The minimum integer k for which there exists a k-coloring of D is called the *chromatic number* $\chi(D)$ of the digraph D. (Note that $\chi(D) \leq |V(D)|$ since D has no loops.) Bokal et al. [4] proved that (in contrast with the undirected case) it is NP-complete to decide whether an input digraph D has $\chi(D) \leq 2$.

Let F be a fixed digraph. An *F-coloring* of a digraph D is a digraph homomorphism $D \to F$, i.e., a mapping $\phi\colon V(D) \to V(F)$ such that $\phi(u)\phi(v)$ is an arc of F whenever uv is an arc of D. The *F-coloring problem* asks whether or not an input digraph D admits an F-coloring [1, 2, 3, 12, 13]. In contrast to the case of graphs, no complexity classification of F-coloring problems is known or conjectured. In fact, it is not even known if each F-coloring problem is polynomial time solvable or NP-complete, and if such a *dichotomy* result were true, then a much more general dichotomy for all constraint satisfaction problems would also hold [6]. There is, however, a conjecture [2] proposing a classification of the complexity of F-coloring problems when each vertex of F has a positive indegree as well as a positive outdegree. Our last result, Theorem 3.1, verifies a special case of this conjecture.

A graph G defines a natural digraph $D(G)$ with the same vertices as G, in which uv is an arc if and only if u and v are adjacent in G. Note that $D(G)$ is a *symmetric* digraph, i.e., the reversal of each arc is an arc. It is easy to see that a mapping $f : V(G) \to V(H)$ is a homomorphism of the graph G to the graph H if and only if it is a homomorphism of the digraph $D(G)$ to the digraph $D(H)$. (We say that the definition of digraph homomorphisms is *consistent* with the definition of graph homomorphisms.)

We introduce a different kind of digraph homomorphism and obtain a complete classification of the corresponding F-coloring problems.

An *acyclic homomorphism* of a digraph D into a digraph F is a mapping $\phi\colon V(D) \to V(F)$ such that

(i) for every arc $uv \in E(D)$, either $\phi(u) = \phi(v)$ or $\phi(u)\phi(v)$ is an arc of F, and
(ii) for every vertex $v \in V(F)$, the subgraph of D induced on $\phi^{-1}(v)$ is acyclic.

It is easy to check that the composition of acyclic homomorphisms is again an acyclic homomorphism. It is also easy to see that this definition is also consistent with the definition of graph homomorphisms, i.e., that a mapping f is a graph homomorphism of G to H if and only if it is an acyclic digraph homomorphism of $D(G)$ to $D(H)$.

An acyclic homomorphism of D to F will also be called an *acyclic F-coloring* of D. For a fixed digraph F, the *acyclic F-coloring problem* asks whether or not an input digraph D admits an acyclic F-coloring.

We now define a digraph analogue of $C(k,d)$: The digraph $\vec{C}(k,d)$ has the vertex set $V(\vec{C}(k,d)) = \{0,\dots,k-1\}$, and from each vertex $i \in V(\vec{C}(k,d))$ there are arcs to $i+d$, $i+d+1$, $\dots$, $i+k-1$, with arithmetic modulo k. Notice that $\vec{C}(n,n-1) \simeq \vec{C}_n$ is the directed n-cycle.

One can again define the *circular chromatic number* $\chi_c(D)$ of the digraph D [4] as the minimum of all rational numbers k/d (where k and $d \leq k$ are positive integers) such that there exists an acyclic homomorphism $D \to \vec{C}(k,d)$. If k and d are positive integers with $k \geq d$, then $\chi_c(\vec{C}(k,d)) = \frac{k}{d}$.

It is not difficult to see [4] that

$$\chi(D) - 1 < \chi_c(D) \leq \chi(D).$$

It follows from [4] that it is NP-complete to decide if $\chi_c(D) \leq 2$. This suggests that deciding if $\chi_c(D) \leq q$ should be NP-complete for every $q \geq 2$, but it gives no insight on what may hold for $q < 2$.

In this paper we verify that the acyclic F-coloring problem is NP-complete, unless F is acyclic, in which case it is polynomial time solvable. This implies, in particular, that to decide if $\chi_c(D) \leq q$ is also NP-complete for every fixed rational number $q > 1$. Refining this proof, we also conclude that certain F-coloring problems are NP-complete, verifying special cases of two conjectures from [1, 2].

2. Acyclic homomorphisms and colorings. We begin by disposing of the easy positive direction.

PROPOSITION 2.1. *Suppose F is an acyclic digraph. Then a digraph D admits an acyclic F-coloring if and only if D is itself acyclic.*

Proof. If D is acyclic, any constant mapping (all vertices of D map to one vertex of F) is an acyclic homomorphism. Conversely, if D contains a directed cycle C, then any acyclic homomorphism of D to a digraph G takes C to a directed cycle in G. □

For the negative results, we observe that all acyclic F-coloring problems are in the class NP, with the mapping ϕ itself being a concise certificate.

Recall that $\vec{C}_2 = \vec{C}(2,1)$ denotes the directed two-cycle. Note that D admits an acyclic homomorphism to $\vec{C}_2$ if and only if $\chi(D) \leq 2$. Therefore our first negative result follows from [4].

PROPOSITION 2.2. *The acyclic $\vec{C}_2$-coloring problem is NP-complete.*

Proof. We shall present a brief proof, slightly adapting the proof in [4], because we shall need to refer to the details of it in the next section. We shall give a polynomial time reduction from the NP-complete problem of 2-colorability of 3-uniform hypergraphs (also known as the not-all-equal 3-satisfiability problem without negated variables). For such a hypergraph X we construct a digraph D consisting of one vertex x for each vertex x of X and three vertices a_e, b_e, c_e for each hyperedge $e = abc$ of X. The arcs of D are xx_e and x_ex for each vertex x of X and each hyperedge e containing x, and a_eb_e, b_ec_e, c_ea_e for each hyperedge e of X. We claim that X is 2-colorable if and only if D admits an acyclic $\vec{C}_2$-coloring, i.e., can be colored with two colors so that each color class is acyclic. Given a 2-coloring of X, we can apply the same colors to the vertices x of D and the opposite color to all vertices x_e for edges e containing x. There will be no monochromatic directed cycle. Moreover, whenever D is colored with two colors without a monochromatic directed cycle, the coloring of the vertices x yields a 2-coloring of X. □

Recall that $D(G)$ is the symmetric digraph associated with the graph G. On the other hand, each digraph D is also associated with a natural graph $H(D)$ which has the same vertices as D, and in which two vertices u, v are adjacent if and only if both uv and vu are arcs of D. Note that the symmetric digraph $D(H(F))$ is obtained from F by removing all arcs uv for which the reversal vu is not an arc. We call this digraph the *symmetric part* of F. It is again easy to see that if a mapping is a digraph homomorphism of D to F, then it is also a graph homomorphism of the symmetric part of D to the symmetric part of G.

Our second negative result follows from [10].

PROPOSITION 2.3. *If the symmetric part of F contains an odd cycle, then the acyclic F-coloring problem is NP-complete.*

Proof. If the symmetric part of F contains an odd cycle, then $H(F)$ is nonbipartite, and hence it is NP-complete to decide if an input graph G admits a homomorphism to $H(F)$ [10]. On the other hand, we claim that G admits a homomorphism to $H(F)$ if and only if $D(G)$ admits an acyclic homomorphism to F. Any homomorphism of G to $H(F)$ is clearly also an acyclic homomorphism of $D(G)$ to F. Thus consider an acyclic homomorphism ϕ of $D(G)$ to F. Since $D(G)$ is a symmetric digraph, ϕ is in fact an acyclic homomorphism of $D(G)$ to the symmetric part of F, i.e., to $D(H(F))$. Therefore ϕ is a homomorphism of G to $H(F)$. □

We are now ready for our first main result.

THEOREM 2.4. *If F contains a directed cycle, then the acyclic F-coloring problem is NP-complete.*

Proof. Let k be the minimum length of a directed cycle in F. We first assume that $k \geq 3$, i.e., that the symmetric part of F is empty. Let F' be the digraph obtained from F by adding an arc uv whenever there is in F a directed path from u to v of length at most $k-1$. Let D' be the digraph obtained from D by replacing each arc xy by a directed path of length $k-1$ from x to y. We claim that there exists an acyclic homomorphism of D to F' if and only if there exists an acyclic homomorphism of D' to F.

Suppose first that ϕ is an acyclic homomorphism of D' to F. Each arc xy of D corresponds to a path of length $k-1$ from x to y in D', which is taken by the acyclic homomorphism ϕ to a path of length at most $k-1$ in F. (This follows from the definition of an acyclic homomorphism and the fact that there are no directed cycles of length less than k in F.) Thus $\phi(x) = \phi(y)$ or $\phi(x)\phi(y)$ is an arc of F'. Moreover, for every $v \in V(F')$, the set $\phi^{-1}(v) \cap V(D)$ is a subset of $\phi^{-1}(v)$ in D' and hence is acyclic in D'. Observe that if $\phi(x) = \phi(y) = v$, then ϕ maps to v all vertices of D' on the $(k-1)$-path from x to y. Therefore, the set $\phi^{-1}(v) \cap V(D)$ induces an acyclic subgraph of D. Thus ϕ restricted to $V(D)$ is an acyclic homomorphism of D to F'. Conversely, suppose that ϕ is an acyclic homomorphism of D to F'. Then it is easy to see that the mapping ϕ can be extended to all vertices $v \in V(D') \setminus V(D)$ (on the added directed paths of length $k-1$) so that the resulting mapping is an acyclic homomorphism of D' to F.

This argument is a polynomial reduction from the problem of acyclic F'-coloring to the problem of acyclic F-coloring. Since F contains a directed cycle of length k, the digraph F' contains $k \geq 3$ vertices in a complete directed digraph, i.e., the symmetric part of F' contains a triangle. By Proposition 2.3 the acyclic F'-coloring problem, and hence also the acyclic F-coloring problem, is NP-complete.

It remains to deal with the case when the symmetric part of F is bipartite but not empty. Suppose $H(F)$ has $\ell \geq 1$ edges. For any digraph D we construct, in polynomial time, a digraph $D^{(\ell)}$ consisting of disjoint copies $D(i,j)$ of D for all pairs $i < j$, with $i, j = 0, 1, \ldots, \ell$, and of special vertices $a_0, a_1, \ldots, a_\ell, b_0, b_1, \ldots, b_\ell$. Moreover, each vertex of $D(i,j)$ has an arc from a_i and b_i, and to a_j and b_j, and there are also arcs $a_i b_i, b_i a_i$ for all $i = 0, 1, \ldots, \ell$ amongst the special vertices (see Figure 2.1).

We claim that D has an acyclic $\vec{C}_2$-coloring if and only if $D^{(\ell)}$ has an acyclic F-coloring. Indeed, if D has an acyclic $\vec{C}_2$-coloring, then all $D(i,j)$ can be acyclically $\vec{C}_2$-colored by the same $\vec{C}_2$, and this coloring extends to the special vertices as well by coloring all a_i with one color and all b_i with the other. Conversely, if $D^{(\ell)}$ has an

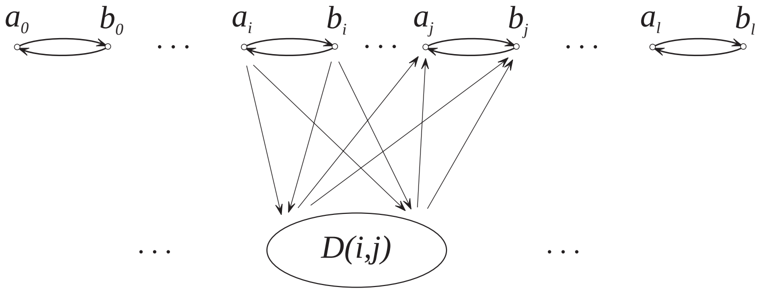


FIG. 2.1. *The digraph $D^{(\ell)}$.*

acyclic F-coloring, then two pairs a_i, b_i and a_j, b_j must map to the same two vertices u, v belonging to an edge of $H(F)$ by the pigeon-hole principle. This means that each vertex c of $D(i,j)$ must also map to u or v, otherwise u, v, and $\phi(c)$ would form a symmetric triangle, contrary to the assumption that $H(F)$ is bipartite. Thus we obtain a $\vec{C}_2$-coloring of D. This amounts to a polynomial reduction of the problem of acyclic $\vec{C}_2$-coloring (which is NP-complete by Proposition 2.2) to the problem of acyclic F-coloring, and hence the latter problem is also NP-complete. □

COROLLARY 2.5. *For every fixed rational number $q > 1$, it is NP-complete to decide if $\chi_c(D) \leq q$.*

Proof. We have $\chi_c(D) \leq \frac{k}{d}$ if and only if D admits an acyclic homomorphism to $\vec{C}(k,d)$, and as long as $d < k$, $\vec{C}(k,d)$ is not acyclic. □

For graphs, it has been shown in [9] that it is NP-hard to decide whether a graph G of chromatic number n satisfies $\chi_c(G) \leq n - \frac{1}{k}$ for any positive integers $k \geq 2$ and $n \geq 3$. One can ask similar questions for circular chromatic numbers of digraphs. We only remark that it is NP-hard to decide if $\chi_c(D') \leq \frac{3}{2}$ even knowing that $\chi(D') = 2$: Consider the digraph $F = \vec{C}_3 = \vec{C}(3,2)$, and apply the proof of Theorem 2.4, with $k = 3$. The digraph F' will be the symmetric triangle, and acyclic F'-colorability is NP-complete. From that proof we know that an input digraph D has an acyclic F'-coloring if and only if the digraph D' (which has chromatic number 2) has an acyclic F-coloring, i.e., has $\chi_c(D') \leq \frac{3}{2}$.

It would also be interesting to know how the complexity of the acyclic F-coloring problem changes when the inputs are restricted. Typical restrictions may involve maximum degree, or planarity, etc. (For undirected graphs we direct the reader to [11] for a survey.) We first make the following observation.

COROLLARY 2.6. *The acyclic $\vec{C}_3$-coloring problem is NP-complete even when restricted to planar digraphs.*

Proof. Let $F = \vec{C}_3$. Since the shortest directed cycle in F has length three, we can apply the above reduction from the problem of acyclic F'-coloring to the problem of acyclic F-coloring. In this case F' is the symmetric triangle; since 3-coloring is NP-complete for planar graphs [7], the corollary follows. □

We also have a similar result for acyclic $\vec{C}_2$-coloring.

THEOREM 2.7. *The acyclic $\vec{C}_2$-coloring problem is NP-complete even when restricted to planar digraphs.*

Proof. We reduce the problem of planar 3-satisfiability. An instance of 3-satisfiability is *planar* if its associated graph is planar. (The associated graph has a vertex C for each clause and a vertex x for each variable; there is an edge joining x and C if variable

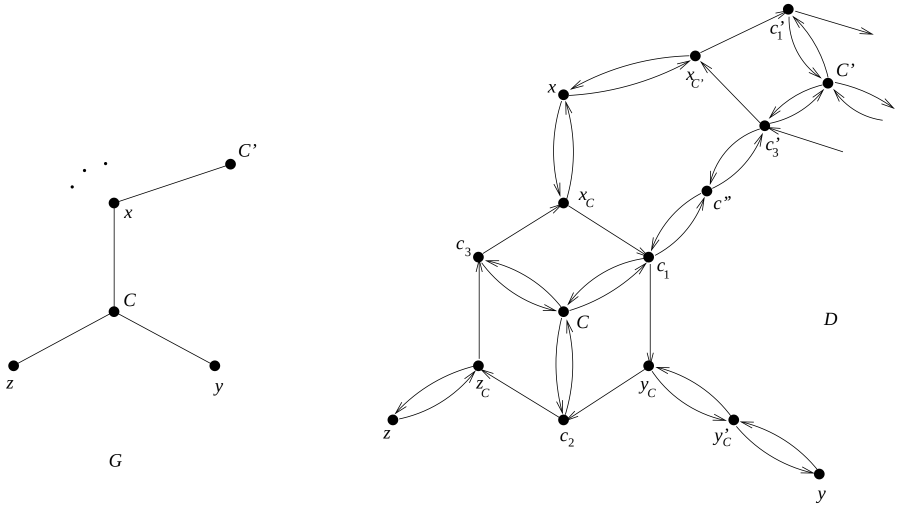


FIG. 2.2. *Construction of D around the clause $C = \neg x \vee y \vee \neg z$.*

x occurs in clause C, positively or negatively.) It is well known that 3-satisfiability is NP-complete even when restricted to connected planar instances [7].

Thus suppose we have an instance of planar 3-satisfiability, and consider the planar embedding of its (connected) associated graph G. We shall transform G to a digraph D which is $\vec{C}_2$-colorable if and only if the instance was satisfiable. The digraph D will contain all the vertices (C and x) of G in the same position in the plane as in G. If C was joined to x, y, z (in this clockwise order) in G, we surround it with a directed six-cycle $x_C c_1 y_C c_2 z_C c_3 x_C$, joined to C by the symmetric set of arcs Cc_1, c_1C, Cc_2, c_2C, Cc_3, c_3C. The new vertices c_i, called *dummy vertices*, are distinct for each clause C. Further, we replace each edge xC of G by the symmetric arcs xx_C, x_Cx if x occurs negatively in C, or the symmetric path of length two xx'_C, x'_Cx, x'_Cx_C, $x_Cx'_C$ if x occurs positively in C. It is clear that the digraph constructed so far is planar. Now consider, for each vertex x corresponding to a variable, the six-cycles corresponding to the clauses C in which x occurs (positively or negatively), in their circular order of the planar embedding. For any two consecutive six-cycles there exist two dummy vertices c, c' which can be joined without destroying the planar embedding; we add the symmetric path of length two $cc'', c''c$, $c''c', c'c''$. This is our planar digraph D. This construction around the clause $C = \neg x \vee y \vee \neg z$ and with consecutive neighbors C and C' around x is represented in Figure 2.2.

We now claim that D admits an acyclic $\vec{C}_2$-coloring if and only if the original instance was satisfiable. Indeed, given a satisfying truth assignment, color each vertex corresponding to a variable x by 0 if x is false and by 1 if x is true, and do the same for all vertices x_C. Furthermore, color all dummy vertices by 0, and color all clause vertices C by 1. It is easy to see that all the auxiliary vertices x'_C and c'' can be colored as well so that the result is an acyclic $\vec{C}_2$-coloring of D. Conversely, suppose we have an acyclic $\vec{C}_2$-coloring of D. Because of the two-cycles Cc_iC, all dummy vertices in any one six-cycle must obtain the same color; because of the symmetric paths of length two between dummy vertices of consecutive six-cycles, all dummy vertices must obtain the same color, say color 0. (Recall that we have assumed that G is connected.) It is now easy to see that the coloring defines a satisfying truth assignment. (Because of the 6-cycles $x_C c_1 y_C c_2 z_C c_3 x_C$, at least one of the vertices x_C, y_C, or z_C has color 1.) □

3. A refinement for *F*-colorings. For a digraph F, let F^p denote the digraph obtained by replacing each vertex of F by the transitive tournament T on $1, 2, \dots, p$. (The arcs of T are all pairs ij with $i < j$.) There is an arc in F^p from a vertex in the copy of T corresponding to u to a vertex in the copy of T corresponding to v if and only if there is an arc from u to v in F. Then it follows from the definitions that a digraph D admits an acyclic homomorphism to F if and only if it admits a homomorphism to F^p with $p = |V(D)|$.

We let similarly F^ω be obtained from F by replacing each vertex by the countable transitive tournament on $1, 2, \dots$. Theorem 2.4 shows that if F is not acyclic, then the F^ω-coloring problem (appropriately defined for mappings of finite digraphs to a finitely described fixed infinite graph) is intractable. We now refine the result to prove that already the F^2-coloring problem is intractable. More precisely, assume for each vertex v of F, we have an integer $p_v \geq 2$. Let F^* be a digraph obtained the same way from F by replacing each v with the transitive tournament on p_v vertices and defining the arcs between these tournaments as above.

THEOREM 3.1. *If F is not acyclic, then the F^*-coloring problem is NP-complete.*

Proof. If the symmetric part of F contains an odd cycle, then the symmetric part of F^* also contains an odd cycle (and we need only each $p_v \geq 1$ here), and the F^*-coloring problem is NP-complete by exactly the same proof as in Proposition 2.3. (Just substitute F^* for F and omit all the occurrences of the word "acyclic.")

If the symmetric part of F is empty, then assume as above that the length of the shortest directed cycle in F is k, where $k \geq 3$. Suppose first that k is odd. Let F' be the digraph on the same vertex set as F^* and with an arc uv whenever there is in F^* a directed path from u to v of length $\frac{k+1}{2}$. A proof similar to the proof of Theorem 2.4 shows that there is a polynomial time reduction from the F'-coloring problem to the F^*-coloring problem. (Take D' to be the digraph obtained from D by replacing each arc xy by a directed path of length $\frac{k+1}{2}$ from x to y. We are using the "indicator construction," Lemma 1 from [10].) We now note that F' contains symmetric pairs of arcs joining vertices at distance $\frac{k+1}{2}$ and $\frac{k-1}{2}$ in the original directed k-cycle in F, and hence the symmetric part of F' contains an odd cycle.

If k is even, we proceed in exactly the same way using directed paths of length $\frac{k}{2} + 1$. In this case the symmetric part of F' also contains a nonbipartite graph when $k \geq 6$. (There are symmetric pairs of arcs joining vertices at distance $\frac{k}{2} - 1$, $\frac{k}{2}$, and $\frac{k}{2} + 1$.) For $k = 4$ we extend our attention to the eight vertices of F^2, a subgraph of F^*, on which the symmetric part of F' is easily seen to have a nonbipartite subgraph. (Indeed, suppose the original 4-cycle in F is $1, 2, 3, 4$, and let $a_1, b_1, a_2, b_2, a_3, b_3, a_4, b_4$ be the corresponding vertices of F^2. Then, using directed paths of length 3, F' contains the symmetric five-cycle $a_1 b_2 b_4 a_2 b_3$.)

It remains to prove that

- if F has a nonempty and bipartite symmetric part, then F^*-coloring is NP-complete.

We proceed by contradiction, assuming that F has a nonempty bipartite symmetric part and that F^*-coloring is not NP-complete. We may assume that S^*-coloring is NP-complete for any proper subgraph S of F which has a nonempty bipartite symmetric part.

This part of the proof uses the "subindicator construction," Lemma 2 of [10]. To review it briefly, in the special case that we shall need, we define a digraph to be a *core* if it does not admit a homomorphism to a proper subgraph. If a digraph F is not a core, then it contains a unique, up to isomorphism, subgraph S which is a core; this subgraph S is called the *core of* F. It is clear that a digraph admits an F-coloring if and only if it admits an S-coloring. (Thus the F-coloring and S-coloring problems are equivalent.) Let J be a fixed digraph with specified vertices v and w. The *subindicator construction*, with respect to J, transforms a given core digraph S, with a specified vertex u, to the subgraph S^- induced by the vertex set V^- defined as follows: Let R be the digraph obtained from the disjoint union of J and S by identifying vertices u and v. Then a vertex x of S belongs to V^- just if there is a homomorphism f of R to S such that $f(y) = y$ for all vertices y of S, and $f(w) = x$. Lemma 2 of [10] gives, for a core S, a polynomial time reduction of the S^--coloring problem to the S-coloring problem.

It follows from our assumptions that F^* is a core, otherwise the core of F^* would be some digraph S^* of a proper subgraph S of F which has a nonempty bipartite symmetric part, and hence both the S^*-coloring and the F^*-coloring problems would be NP-complete.

We first claim that every vertex of F is incident with an edge of $H(F)$. Otherwise, consider the subindicator J consisting of three vertices v, w, and z and two arcs wz, zw. If F contained a vertex x which is not incident with an edge of $H(F)$, then all the p_x vertices of F^* in the transitive tournament replacing x would be missing from $(F^*)^-$ (the vertex u of F^* can be chosen arbitrarily). Thus $(F^*)^-$ would be some S^* where S is a proper subgraph of F which has a nonempty bipartite symmetric part, and hence again both the S^*-coloring and the F^*-coloring problems would be NP-complete.

Next we claim that F cannot have a vertex a and arcs ax, ay such that xy is an edge of $H(F)$. If this were the case, then consider the subindicator J consisting of two vertices v and w and the arc vw, and let u be the last vertex in the transitive tournament of F^* replacing the vertex a of F. Then the digraph $(F^*)^-$ is missing all the p_a vertices of the tournament replacing a but contains the symmetric pairs of arcs arising from x and y. Hence $(F^*)^-$ is some S^* where S is a proper subgraph of F which has a nonempty bipartite symmetric part, and we obtain a contradiction as before.

Finally, we claim that F is a symmetric digraph. Otherwise there would be an arc ab in F such that ba is not an arc of F. Consider the subindicator J consisting of three vertices v, s, w and three arcs vs, sw, ws, and let u be the first vertex in the transitive tournament of F^* replacing the vertex a of F. We first observe that all the p_b vertices of F^* replacing b are missing from $(F^*)^-$: Indeed, since there are no arcs from these p_b vertices to the p_a vertices of F^* replacing a, the only way the vertex w of R can map to one of these p_b vertices, say vertex y, is if there are in F some arcs ux, xy, yx, contradicting the preceding claim. Now we recall that each vertex of F is incident with an edge of $H(F)$; thus there are in F some arcs ac, ca. It follows that $(F^*)^-$ contains the symmetric pairs of arcs arising from the tournaments replacing a and c. This once again contradicts the minimality of F.

Since F is a bipartite symmetric digraph, the core of F must be $\vec{C}_2$, and we need only to consider $F = \vec{C}_2$. In this case F^*-coloring is NP-complete by the same argument as given in the proof of Proposition 2.2. One needs only to note that in any coloring of the digraph D with two colors, each monochromatic set of vertices is not only acyclic: it is a disjoint union of isolated arcs. This means that F^*, with its at least two vertices in a transitive tournament replacing each vertex, has the property that the hypergraph X admits a 2-coloring if and only if the digraph D has an F^*-coloring. Therefore F^*-coloring is NP-complete. □

This result verifies a special case of Conjecture 5.1 in [1] and of Conjecture 6.1 in [2]. In particular, Conjecture 6.1 of [2] states that, for connected digraphs F which have all indegrees and all outdegrees at least one, F-coloring is NP-complete unless the core of F is $\vec{C}_k$ for some integer k (in which case it is known to be polynomial time solvable).

Note that we do not know what the complexity of F^*-coloring is when F is acyclic. Certainly, the problem can be polynomial time solvable: For instance, if F is a transitive tournament, then F^* is also a transitive tournament, and so D is F^*-colorable if and only if it is acyclic and has height no greater than $|V(F^*)|$ (the height of F^*). Similarly, the problem can be NP-complete: For instance, there are acyclic triangle-free digraphs F (even oriented trees F [12]) such that F-coloring is NP-complete. Then F^p-coloring is also NP-complete, since an input digraph D is F-colorable if and only if D^p is F^p-colorable. One only needs to notice that the fact that F is triangle-free implies that the $2p$-vertex tournaments of D^p corresponding to edges of D must map to the $2p$-vertex tournaments of F^p corresponding to the edges of F.

REFERENCES

[1] J. Bang-Jensen, P. Hell, and G. MacGillivray, *On the complexity of colouring by supergraphs of bipartite graphs*, Discrete Math., 109 (1992), pp. 27–44.

[2] J. Bang-Jensen, P. Hell, and G. MacGillivray, *The effect of two cycles on the complexity of colourings by directed graphs*, Discrete Appl. Math., 26 (1990), pp. 1–23.

[3] J. Bang-Jensen, P. Hell, and G. MacGillivray, *The complexity of colourings by semicomplete digraphs*, SIAM J. Discrete Math., 1 (1988), pp. 281–298.

[4] D. Bokal, G. Fijavž, P. M. Kayll, M. Juvan, and B. Mohar, *The circular chromatic number of a digraph*, J. Graph Theory, submitted.

[5] J. A. Bondy and P. Hell, *A note on the star chromatic number*, J. Graph Theory, 14 (1990), pp. 479–482.

[6] T. Feder and M. Y. Vardi, *The computational structure of monotone monadic SNP and constraint satisfaction: A study through Datalog and group theory*, SIAM J. Comput., 28 (1998), pp. 57–104.

[7] M. R. Garey and D. S. Johnson, *Computers and Intractability, A Guide to the Theory of NP-completeness*, W. H. Freeman, San Francisco, 1979.

[8] D. R. Guichard, *Acyclic graph coloring and the complexity of the star chromatic number*, J. Graph Theory, 17 (1993), pp. 129–134.

[9] H. Hatami and R. Tusserkani, *On the Complexity of the Circular Chromatic Number*, manuscript, 2002.

[10] P. Hell and J. Nešetřil, *On the complexity of H-coloring*, J. Combin. Theory Ser. B, 48 (1990), pp. 92–110.

[11] P. Hell and J. Nešetřil, *Counting list homomorphisms and graphs with bounded degrees*, Discrete Math., to appear.

[12] P. Hell, J. Nešetřil, and X. Zhu, *Duality and polynomial testing of tree homomorphisms*, Trans. Amer. Math. Soc., 348 (1996), pp. 147–156.

[13] H. A. Maurer, J. H. Sudborough, and E. Welzl, *On the complexity of the general colouring problem*, Inform. and Control, 51 (1981), pp. 123–145.

[14] B. Mohar, *Circular colorings of edge-weighted graphs*, J. Graph Theory, 43 (2003), pp. 107–116.

[15] X. Zhu, *Circular chromatic number: A survey*, Discrete Math., 229 (2001), pp. 371–410.

SIAM J. DISCRETE MATH.
Vol. 17, No. 2, pp. 171–195

BOUNDS ON THE LIST-DECODING RADIUS OF REED–SOLOMON CODES[*]

GITIT RUCKENSTEIN[†] AND RON M. ROTH[†]

Abstract. Techniques are presented for computing upper and lower bounds on the number of errors that can be corrected by list decoders for general block codes and, specifically, for Reed–Solomon (RS) codes. The list decoder of Guruswami and Sudan implies such a lower bound (referred to here as the GS bound) for RS codes. It is shown that this lower bound, given by means of the code's length, the minimum Hamming distance, and the maximal allowed list size, in fact applies to all block codes. Ranges of code parameters are identified where the GS bound is tight for worst-case RS codes, in which case the list decoder of Guruswami and Sudan provably corrects the largest possible number of errors.

On the other hand, ranges of parameters are provided for which the GS lower bound can be strictly improved. In some cases the improvement applies to all block codes with a given minimum Hamming distance, while in others it applies only to RS codes.

Key words. block designs, Guruswami–Sudan algorithm, list decoding, MDS codes, Reed–Solomon codes, Sidon sets

AMS subject classifications. 11T71, 05B05

DOI. 10.1137/S0895480101395506

1. Introduction.

An (n, M, d) (block) code $\mathcal{C}$ over an alphabet F of size q is an M-subset of F^n with minimum Hamming distance d between any two different codewords. In cases where F is a finite field and $\mathcal{C}$ is a linear subspace of F^n, namely, $k = \log_q M = \dim \mathcal{C}$, we refer to $\mathcal{C}$ as an $[n, k, d]$ code. An (n, M, d) code is called maximum-distance separable (MDS) [11, Ch. 11] if $d = n+1-\log_q M$, thus satisfying the Singleton bound [3, p. 88] with equality; in particular, $k = \log_q M$ must be an integer.

An $[n, k, d]$ *(generalized) Reed–Solomon (RS) code* over a finite field $F = \mathrm{GF}(q)$ is a linear MDS code that consists of all words (vectors) of the form $(f(\alpha_1)\ f(\alpha_2)\ \cdots\ f(\alpha_n))$, where $\alpha_1, \alpha_2, \ldots, \alpha_n$ are prescribed distinct elements of F, which are commonly referred to as the *code locators*, and $f(x)$ ranges over all polynomials of degree less than $k = n-d+1$ over F.

Denote by $\mathsf{d}_H(\boldsymbol{v}_1, \boldsymbol{v}_2)$ the Hamming distance between two words $\boldsymbol{v}_1, \boldsymbol{v}_2 \in F^n$. A *list-$\ell$ decoder with a decoding radius* τ for a code $\mathcal{C} \subseteq F^n$ is a mapping $\mathcal{D} : F^n \longrightarrow 2^{\mathcal{C}}$ such that (i) $|\mathcal{D}(\boldsymbol{v})| \leq \ell$ for every $\boldsymbol{v} \in F^n$, and (ii) $\boldsymbol{c} \in \mathcal{D}(\boldsymbol{v})$ if and only if $\boldsymbol{c} \in \mathcal{C}$ and $\mathsf{d}_H(\boldsymbol{c}, \boldsymbol{v}) \leq \tau$. In other words, given a received word $\boldsymbol{v} \in F^n$, the decoder $\mathcal{D}$ returns all the codewords in $\mathcal{C}$ that are at Hamming distance at most τ from $\boldsymbol{v}$, and the size of that list is guaranteed to be at most ℓ. The decoding radius τ therefore stands for the largest number of errors that are corrected by $\mathcal{D}$.

Denote by $\Delta_\ell(\mathcal{C})$ the largest decoding radius of any list-ℓ decoder for a code $\mathcal{C} \subseteq F^n$. The value $\Delta_\ell(\mathcal{C})$ is the largest integer value R such that all Hamming spheres of radius R in F^n contain at most ℓ codewords of $\mathcal{C}$.

[*] Received by the editors September 21, 2001; accepted for publication (in revised form) May 5, 2003; published electronically November 4, 2003. This work was supported by grant 94/99 from the Israel Science Foundation. This work was presented in part at the 2001 International Symposium on Information Theory (ISIT 2001), Washington, D.C.

http://www.siam.org/journals/sidma/17-2/39550.html

[†] Computer Science Department, Technion, Haifa 32000, Israel (gitit@cs.technion.ac.il, ronny@cs.technion.ac.il).

Hereafter, by an *admissible quadruple* (ℓ, n, d, q), we mean that ℓ, n, d, and q are positive integers such that $1 \le d \le n$. By an *RS-admissible quadruple* (ℓ, n, d, q), we mean an admissible quadruple for which, in addition, $n \le q$ and q is a power of a prime.

Given an admissible quadruple (ℓ, n, d, q), we define

$$\Delta_\ell(n, d; q) = \min_{\mathcal{C}} \Delta_\ell(\mathcal{C}) \ , \tag{1}$$

where the minimum is taken over all (n, M, d) block codes over an alphabet of size q. For an RS-admissible quadruple (ℓ, n, d, q), we also define

$$\Delta_\ell^{\mathrm{RS}}(n, d; q) = \min_{\mathcal{C}} \Delta_\ell(\mathcal{C}) \ , \tag{2}$$

where the minimum is taken over all $[n, k, d]$ RS codes over $\mathrm{GF}(q)$. Studying these two quantities is the subject of this paper. Taking the minimum in (1) or (2) results in the value $\Delta_\ell(\mathcal{C})$ of the "worst" code $\mathcal{C}$ in the respective family. In particular, we are interested here in the attainable performance of list-ℓ decoders of RS codes (i.e., in the largest number of errors that can be corrected by such decoders), independent of the particular choice of the code locators. From the practical side, this is justified by the structure of existing RS decoding algorithms, which are typically not tailored to specific selection of code locators. When $n = q$, the minimum in the definition of $\Delta_\ell^{\mathrm{RS}}(n, d; q)$ is taken over one set of code locators; in fact, this is also the case when $n = q-1$, where one can assume that all the code locators are nonzero (see [11, p. 305, Problem 7]).

Clearly, the quantities $\Delta_\ell(n, d; q^m)$ and $\Delta_\ell^{\mathrm{RS}}(n, d; q^m)$ are nondecreasing with ℓ and nonincreasing with m, and for every admissible quadruple (ℓ, n, d, q),

$$\Delta_\ell(n, d; q) \le \Delta_\ell^{\mathrm{RS}}(n, d; q) \, .$$

It is well known that

$$\Delta_1(n, d; q) = \Delta_1^{\mathrm{RS}}(n, d; q) = \lfloor (d-1)/2 \rfloor$$

(independent of q).

1.1. The Guruswami–Sudan bound. Guruswami and Sudan present in [7] a list-ℓ decoding algorithm for $[n, k, d]$ RS codes over $\mathrm{GF}(q)$ (see also the earlier work of Sudan [13]). The decoding radius of their decoder depends on the parameters (ℓ, n, d, q), as summarized in Theorem 1.1 below. We first introduce several notations that are required not only for the statement of their result, but also in our analysis throughout this paper.

Given $\ell \ge 1$, partition the real interval $[0, 1)$ into the ℓ subintervals

$$[0, \rho_2), [\rho_2, \rho_3), \ldots, [\rho_\ell, 1) \ , \tag{3}$$

where

$$\rho_r = \rho_r(\ell) = \frac{r(r-1)}{\ell(\ell+1)} \ , \quad r = 1, 2, \ldots, \ell+1 \, . \tag{4}$$

Given integers n and d such that $1 \le d \le n$, define the relative minimum distance $\delta = d/n$. Let $r = r(\delta)$ be the unique index such that $1 - \delta \in [\rho_r, \rho_{r+1})$. Also define

$$\tau_\ell(n, d) = \frac{1}{(\ell + 1)r} \left(\tbinom{\ell+1}{2} d - \tbinom{\ell+1-r}{2} n \right) \, . \tag{5}$$

The mapping $\delta \mapsto \tau_\ell(n, n\delta)$ is piecewise-linear and continuous over $[0,1)$ for every fixed n. It can be easily verified that $\tau_\ell(n,d) < d$ for every value of ℓ, assuming $d \leq n$. One can also verify that, when $1 - \delta \geq \rho_\ell$,

$$\tau_\ell(n,d) = d/2 \; .$$

By its definition, $\tau_\ell(n,d)$ is an integer if and only if

$$(\ell{+}1)r \quad \text{divides} \quad \tbinom{\ell+1}{2}d - \tbinom{\ell+1-r}{2}n \; . \tag{6}$$

The following result follows from [7] and is proved in the appendix.

THEOREM 1.1. *For every RS-admissible quadruple (ℓ, n, d, q), the list-ℓ decoder in* [7] *for an $[n,k,d]$ RS code over* GF(q) *has a decoding radius $\lceil \tau_\ell(n,d) \rceil - 1$; so,*

$$\Delta_\ell^{RS}(n,d;q) \geq \lceil \tau_\ell(n,d) \rceil - 1 \; . \tag{7}$$

1.2. Main results. In this paper, we first generalize the result of [7] by showing that $\lceil \tau_\ell(n,d) \rceil - 1$, referred to as the *GS bound*, is a lower bound on the list-ℓ decoding radius of every (n, M, d) block code, as stated in Theorem 1.2.

THEOREM 1.2. *Let (ℓ, n, d, q) be an admissible quadruple. Then*

$$\Delta_\ell(n,d;q) \geq \lceil \tau_\ell(n,d) \rceil - 1 \; . \tag{8}$$

Theorem 1.2, which is similar to a result by Johnson [11, Ch. 17, Thm. 2], is proved in section 2.1 by using combinatorial arguments, while the result in Theorem 1.1 is based on algebraic analysis. When $d/n \leq 2/(\ell{+}1)$ $(= 1 - \rho_\ell)$, (8) becomes

$$\Delta_\ell(n,d;q) \geq \lfloor (d{-}1)/2 \rfloor = \Delta_1(n,d;q) \; .$$

In a recent work [9], Justesen and Høholdt compute RS-admissible quadruples (ℓ, n, d, q) for which there exist (n, q^{n-d+1}, d) MDS and RS codes over $F = \mathrm{GF}(q)$ that attain the GS bound. A key ingredient in their technique is constructing what we call here a *failing list* of codewords. By a failing list of size $\ell{+}1$, we mean a set of $\ell{+}1$ words, $\{\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_\ell\} \subseteq F^n$, such that the following two conditions hold:

- $\mathsf{d}_H(\boldsymbol{c}_s, \boldsymbol{c}_t) \geq d$ for every $0 \leq s < t \leq \ell$, and
- there is some $\boldsymbol{v} \in F^n$ such that $\mathsf{d}_H(\boldsymbol{c}_s, \boldsymbol{v}) \leq \lceil \tau_\ell(n,d) \rceil$ for every $0 \leq s \leq \ell$.

One can easily see that a failing list of size $\ell{+}1$ is contained in an (n, M, d) code $\mathcal{C}$ if and only if $\Delta_\ell(\mathcal{C})$ attains the GS bound. Several families of MDS codes and RS codes that contain such failing lists are presented in [9]; their constructions are based on block designs, and in each of these constructions, the relative minimum distance δ is such that $1{-}\delta = \rho_r(\ell)$.

In this work, we introduce a combinatorial configuration, akin to block designs, that defines a structure of failing lists which covers the *whole* range of rational δ values (and not just those for which $1 - \delta = \rho_r(\ell)$). Furthermore, we prove that for triples (ℓ, n, d) that satisfy the divisibility condition (6), our structure completely characterizes the failing lists of size $\ell{+}1$ in any given (n, M, d) code over any alphabet F. This, in turn, provides sufficient and *necessary* conditions on the existence of such failing lists (see Proposition 2.3 in section 2).

It turns out that our necessary conditions imply that there is a range of parameters where the GS bound is *not* tight for any code. For example, Proposition 1.3 below indicates the nonexistence of failing lists in cases where the alphabet size is small.

Proposition 1.3. *Let (ℓ, n, d, q) be an admissible quadruple, let r be the unique integer such that $1 - d/n \in [\rho_r, \rho_{r+1})$, and assume that* (6) *holds. Then $\Delta_\ell(n, d; q) \geq \tau_\ell(n, d)$ if either of the following conditions holds:*

- *$1 - d/n = \rho_r$ and $q < \ell{+}1{-}r$, or*
- *$1 - d/n > \rho_r$ and $q < \ell{+}2{-}r$.*

Proposition 1.3 is proved in section 2.4, where additional cases are indicated in which the GS bound is not tight. These cases are found by connecting the (non)existence of failing lists to the (non)existence of constant-weight codes and of block designs. (In contrast, Justesen and Høholdt identify triples (ℓ, n, d) for which the GS bound is tight for MDS codes over sufficiently large fields; see (the proof of) Theorem 4 in [9].)

The remaining results in our paper deal with RS codes. Here, we use the identity $k{-}1 = n{-}d$, and we slightly modify the common definition of *rate* of an $[n, k, d]$ MDS code and use it for the value $(k{-}1)/n = 1 - \delta$; as it turns out, this value fits more conveniently into our analysis. The intervals $[\rho_r, \rho_{r+1})$ are thus referred to as rate intervals.

First, we obtain sufficient and necessary conditions for the existence of failing lists in RS codes (see Lemma 3.1 in section 3). Using our sufficient conditions, we identify families of RS codes (other than those obtained in [9]) that attain the GS bound. For triples (ℓ, n, k) that correspond to the first and last subintervals in (3) (specifically, $(k{-}1)/n \leq 2/(\ell(\ell{+}1))$ or $(k{-}1)/n \geq 1 - (2/(\ell{+}1))$), we find a variety of finite fields $\mathrm{GF}(q)$ over which there are $[n, k, d]$ RS codes that attain the GS bound. These results are summarized in Propositions 1.4 and 1.5 below and are proved later on (with all subsequent results that are stated in this section) in section 4.

Proposition 1.4 covers the high-rate range (i.e., small values of d/n) and identifies quadruples (ℓ, n, d, q) for which a list-ℓ decoder for the worst $[n, k, d]$ RS code, and hence for the worst (n, M, d) code, does no better than a list-1 ("classical") decoder.

Proposition 1.4. *Let the RS-admissible quadruple (ℓ, n, d, q), other than $(3, 2, 1, 2)$, satisfy*

$$d/n \;\leq\; \frac{2}{\ell{+}1} \;\left(= 1 - \rho_\ell\right) .$$

Assume in addition that when $d > 1$, the integer $\lceil (d{-}1)/2 \rceil$ divides either $q{-}1$ or q. Then

$$\Delta_\ell^{RS}(n, d; q) = \lceil \tau_\ell(n, d) \rceil - 1 = \lfloor (d{-}1)/2 \rfloor = \Delta_1^{RS}(n, d; q) .$$

We show in section 4.3 that there are infinitely many RS-admissible quadruples that satisfy the conditions of Proposition 1.4.

Proposition 1.5 covers the low-rate range (the leftmost subinterval in (3), namely, high values of d/n) and makes use of the following definition. A subset X of an Abelian group is called a *weak Sidon set* if every four distinct elements $\theta_1, \theta_2, \theta_3, \theta_4 \in X$ satisfy $\theta_1 + \theta_2 \neq \theta_3 + \theta_4$ (see [1], [4], [6], [12]). The notation $\mathbb{Z}_m$ will stand for the ring of integers modulo m.

Proposition 1.5. *For a prime p, let the RS-admissible quadruple $(\ell, n, d, q{=}p^h)$ satisfy*

$$\left(1 - \rho_2 =\right) 1 - \frac{2}{\ell(\ell{+}1)} \;\leq\; d/n \;<\; 1 .$$

Assume in addition that either

(a) $n{-}d \,|\, q{-}1$ *and the additive group of* $\mathbb{Z}_{(q-1)/(n-d)}$ *contains a weak Sidon set of size* $\ell{+}1$*, or*

(b) $n{-}d = p^b$ *for some integer* b *and* $\mathbb{Z}_p^{h-b}$ *contains a weak Sidon set of size* $\ell{+}1$*.*

Then

$$\Delta_\ell^{RS}(n,d;q) = \lceil \tau_\ell(n,d) \rceil - 1 = \lceil \ell n/(\ell{+}1) - \ell(n{-}d)/2 \rceil - 1 \;.$$

Based on known properties of Sidon sets, we show in section 4.5 that each of the two cases, (a) and (b), in Proposition 1.5 covers infinitely many RS-admissible quadruples.

Observe that we have excluded the case $d = n$ (the repetition code) from Proposition 1.5. Here we have

$$\Delta_\ell^{\mathrm{RS}}(n,n;q) = \lceil \tau_\ell(n,n) \rceil - 1 = \lceil (\ell n/(\ell{+}1)) \rceil - 1$$

only when $\ell < q$; there are $\ell{+}1$ codewords at Hamming distance $\leq \lceil \ell n/(\ell{+}1) \rceil$ from a word $\boldsymbol{v}$ in which each of some $\ell{+}1$ elements of $\mathrm{GF}(q)$ occurs at least $\lfloor n/(\ell{+}1) \rfloor$ times. When $\ell \geq q$ we obviously have $\Delta_\ell^{\mathrm{RS}}(n,n;q) = n$.

Consider now the intermediate subintervals in (3), i.e., the midrate range

$$\frac{2}{\ell{+}1} \;<\; \frac{d}{n} \;<\; 1 - \frac{2}{\ell(\ell{+}1)}\;;$$

this range is nonempty for $\ell \geq 3$. The treatment of this range seems to be more elaborate than the extreme (rightmost and leftmost) subintervals. Hence, our results for the midrate range are quite partial; yet, they demonstrate that the techniques that are developed in this paper are applicable not only to the extreme subintervals. These results are presented in section 4.4.

The propositions presented in this introduction section, together with those presented in sections 4.1 and 4.4, imply, for example, that

$$\liminf_{q \to \infty} \Delta_3(n,k;q) = \lceil \tau_3(n,k) \rceil - 1$$

for all $1 \leq k \leq n \leq 15$, except possibly for $(n,k) \in \{(4,2),(10,3),(14,6),(15,7)\}$. Verifying this statement is left to the reader.

On the other hand, as part of our treatment of the midrate range, we also find RS-admissible quadruples (ℓ,n,d,q) for which the GS lower bound is not tight. The next two propositions provide two examples of such quadruples.

PROPOSITION 1.6. *Let* $q \geq 11$ *be a power of an* odd *prime. Then,*

$$\Delta_4^{RS}(10,7;q) \geq \tau_4(10,7) = 4\;.$$

In contrast, we show in section 4.4 that $\Delta_4^{\mathrm{RS}}(10,7;q) = \tau_4(10,7) - 1 = 3$ when q is even. Moreover, it follows from Theorem 4 in [9] that $\Delta_4(10,7;q) = \tau_4(10,7) - 1 = 3$ for every large enough field size q.

PROPOSITION 1.7. *For every* $h \geq 4$*,*

$$\Delta_{10}^{RS}(11,9;2^h) \geq \tau_{10}(11,9) = 6\;.$$

This work is organized as follows. In section 2, we develop the tools for synthesizing and analyzing failing lists in general codes. Theorem 1.2, Proposition 1.3, and some other combinatorial conditions on the tightness of the GS bound are proved using these tools. Specific tools for RS codes are then introduced in section 3. Finally, section 4 contains the proofs for Propositions 1.4–1.7.

2. Failing lists in general codes.

Throughout this section, we fix the alphabet F of size q, the length n and minimum Hamming distance $d < n$ of an (n, M, d) code over F, and a list size ℓ. We let r be the unique integer such that $1 - d/n \in [\rho_r, \rho_{r+1})$, and we use the notation $\langle n \rangle$ for the set $\{1, 2, \ldots, n\}$.

2.1. Lower bound on $\boldsymbol{\Delta_\ell(n, d\,; q)}$.

Proof of Theorem 1.2. Assume to the contrary that there is an (n, M, d) code $\mathcal{C}$ for which $\Delta_\ell(\mathcal{C}) < \lceil \tau_\ell(n,d) \rceil - 1$. It follows that there is a set of $\ell{+}1$ codewords, $\mathcal{L} = \{\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_\ell\} \subseteq \mathcal{C}$, and a word $\boldsymbol{v} \in F^n$ such that $\mathsf{d}_H(\boldsymbol{c}_s, \boldsymbol{v}) \leq \lceil \tau_\ell(n,d) \rceil$ for every $0 \leq s \leq \ell$.

For every $\mu \in \langle n \rangle$, denote by x_μ the number of words in $\mathcal{L}$ that agree with $\boldsymbol{v}$ on the μth position. On the one hand, it is clear that

$$\sum_{\mu=1}^{n} x_\mu > (\ell{+}1)(n - \tau_\ell(n,d)) \,. \tag{9}$$

On the other hand, the number of different (unordered) pairs $\{\boldsymbol{c}_s, \boldsymbol{c}_t\} \subseteq \mathcal{L}$ that agree on their μth coordinate is at least $\binom{x_\mu}{2}$. Since $\mathsf{d}_H(\boldsymbol{c}_s, \boldsymbol{c}_t) \geq d$ for every $0 \leq s < t \leq \ell$, it follows that the total number of agreement coordinates, when ranging over all pairs $\{\boldsymbol{c}_s, \boldsymbol{c}_t\} \subseteq \mathcal{L}$, cannot exceed $\binom{\ell+1}{2}(n{-}d)$; therefore,

$$\sum_{\mu=1}^{n} \binom{x_\mu}{2} \leq \binom{\ell+1}{2}(n{-}d) \,. \tag{10}$$

Define

$$y = \tfrac{1}{r}\left(\binom{\ell+1}{2}(n{-}d) - \binom{r}{2}n\right) \,. \tag{11}$$

By the definition of the parameter r, we get that $0 \leq y < n$. It can be easily verified that the right-hand side of (9) equals $rn{+}y$. Denote $t = \lfloor y \rfloor{+}1$. It follows from (9) that

$$\sum_{\mu=1}^{n} x_\mu \geq rn + t \,. \tag{12}$$

Regard $x_1, x_2, \ldots, x_n$ as integer variables that are constrained to satisfy (12). By [11, p. 526], the minimum of the sum $\sum_{\mu=1}^{n} \binom{x_\mu}{2}$ is

$$\tfrac{1}{2}\big(t(r{+}1)^2 + (n{-}t)r^2 - (rn{+}t)\big) \;=\; \binom{r}{2}n + rt \;>\; \binom{r}{2}n + ry \;=\; \binom{\ell+1}{2}(n{-}d) \,,$$

contradicting (10). □

The above proof is essentially a generalization of the proofs of Theorems 2 and 3 in [11, Ch. 17] to any finite alphabet; Theorem 2 therein is the Johnson bound on the size of a binary constant-weight code, and Theorem 3 follows from the Johnson bound by using arguments that take into account that the parameters we optimize over are integers. It turns out that a similar proof technique can be applied in our case, where nonbinary codes are considered. (In [5, Thm. 4.2 (part 2)], the proof technique of the Johnson bound is refined for nonbinary codes. It uses the observation that two codewords $\boldsymbol{c}_s$ and $\boldsymbol{c}_t$ in a nonbinary code can both disagree with a given word $\boldsymbol{v}$ on a given position i while they also disagree with each other on position i. However, the proof in [5] does not take into account that some of the parameters involved are integer-valued. A further improvement on both Theorem 1.2 and [5, Thm. 4.2 (part 2)] has been recently reported in [14].)

2.2. (ℓ, r)-configurations. We define an (ℓ, r)-*configuration* as a set $\mathcal{L}$ of $\ell+1$ words in F^n such that for every position $\mu \in \langle n \rangle$ the following holds: there are exactly r or $r+1$ words in $\mathcal{L}$ that agree on that position by taking the same value, c_μ, and the remaining $\ell+1-r$ (respectively, $\ell-r$) words are all distinct on that position; neither does any of them take there the value c_μ.

We now repeat the definition with slightly more detail. Let $S_1, S_2, \ldots, S_{\binom{\ell+1}{r}}$ be all the distinct subsets of $\{0, 1, \ldots, \ell\}$ of size r, and let $S'_1, S'_2, \ldots, S'_{\binom{\ell+1}{r+1}}$ be all the distinct subsets of $\{0, 1, \ldots, \ell\}$ of size $r+1$. A *partition vector* of $\langle n \rangle$ is an (ordered) list of $\binom{\ell+2}{r+1} = \binom{\ell+1}{r} + \binom{\ell+1}{r+1}$ disjoint subsets,

$$\left(I_1, I_2, \ldots, I_{\binom{\ell+1}{r}}, I'_1, I'_2, \ldots, I'_{\binom{\ell+1}{r+1}} \right) ,$$

whose union is $\langle n \rangle$. A partition vector $\mathcal{P}$ is said to be *proper* if $I'_j = \emptyset$ for all j. The existence of a proper (ℓ, r)-configuration over F implies $q \geq \ell+1-r$, where the existence of a nonproper configuration implies the weaker inequality $q \geq \ell+2-r$.

We will hereafter abbreviate notation and write $(I_i)_i \| (I'_j)_j$ for a partition vector; a proper partition vector will also be written as $(I_i)_i$. Given a partition vector $\mathcal{P} = (I_i)_i \| (I'_j)_j$, an (ℓ, r)-configuration with respect to $\mathcal{P}$ is a set of $\ell+1$ words $\mathcal{L} = \{\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_\ell\} \subseteq F^n$ that satisfies the following two conditions:

- For every $i = 1, 2, \ldots, \binom{\ell+1}{r}$, the words in $\mathcal{L}_i = \{\boldsymbol{c}_s\}_{s \in S_i}$ are identical on the positions indexed by I_i, while none of the words in $\mathcal{L} \setminus \mathcal{L}_i$ agrees on any of those positions with any other word in $\mathcal{L}$.
- The same as the previous condition, with I'_j replacing I_i and $\mathcal{L}'_j = \{\boldsymbol{c}_s\}_{s \in S'_j}$ replacing $\mathcal{L}_i$ for $j = 1, 2, \ldots, \binom{\ell+1}{r+1}$.

The existence of an (ℓ, r)-configuration $\mathcal{L}$ with respect to a partition vector $\mathcal{P} = (I_i)_i \| (I'_j)_j$ implies the existence of an *incidence structure* $\boldsymbol{D}(\mathcal{L}) = (\mathcal{L}, \mathcal{B}, \mathcal{M})$ (see [2, Ch. 1]) with $\ell+1$ "points," corresponding to the codewords in $\mathcal{L}$, and a multiset $\mathcal{B}$ of n (not necessarily distinct) "blocks." The $(\ell+1) \times n$ incidence matrix $\mathcal{M}$, which represents the incidence relation, is defined as follows:

$$\mathcal{M}_{s,\mu} = \begin{cases} 1 & \text{if } \mu \in I_i \text{ and } s \in S_i \text{ for some } i, \\ 1 & \text{if } \mu \in I'_j \text{ and } s \in S'_j \text{ for some } j, \\ 0 & \text{otherwise.} \end{cases} \qquad s \in \{0, 1, \ldots, \ell\}, \quad \mu \in \langle n \rangle,$$

Using the terminology of [2], $\boldsymbol{D}(\mathcal{L})$ is an incidence structure with possibly repeated blocks and up to two block sizes, r and $r+1$; when the partition elements I_i and I_j are all of size ≤ 1, no repeated blocks appear, and when $\mathcal{P}$ is a proper partition vector, only the block size r is allowed.

Lemma 2.1 below provides sufficient conditions for an (ℓ, r)-configuration $\mathcal{L}$ to form a failing list.

LEMMA 2.1. *Let $\mathcal{L} = \{\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_\ell\}$ be an (ℓ, r)-configuration with respect to a partition vector $(I_i)_i \| (I'_j)_j$ of $\langle n \rangle$ that satisfies both*

$$\sum_{i \,:\, \{s,t\} \subseteq S_i} |I_i| \;+ \sum_{j \,:\, \{s,t\} \subseteq S'_j} |I'_j| \;\leq\; n-d \,, \qquad 0 \leq s < t \leq \ell \,, \tag{13}$$

and

$$\sum_{i \,:\, s \in S_i} |I_i| \;+\; \sum_{j \,:\, s \in S'_j} |I'_j| \;\geq\; n - \lceil \tau_\ell(n, d) \rceil \,, \qquad 0 \leq s \leq \ell \,. \tag{14}$$

Then $\mathcal{L}$ is a failing list; each word in $\mathcal{L}$ is at Hamming distance at most $\lceil \tau_\ell(n,d) \rceil$ from the majority-vote word $\boldsymbol{v} \in F^n$ that agrees on any position in I_i (respectively, I'_j) with the words in $\mathcal{L}_i$ (respectively, $\mathcal{L}'_j$).

Proof. By (13), every two words in $\mathcal{L}$ agree on at most $n{-}d$ positions and, thus, $\mathsf{d}_H(\boldsymbol{c}_s, \boldsymbol{c}_t) \geq d$ for every $0 \leq s < t \leq \ell$. In addition, by (14), the Hamming distance of each word in $\mathcal{L}$ from the word $\boldsymbol{v}$ is not greater than $\lceil \tau_\ell(n,d) \rceil$. It follows that $\mathcal{L}$ is a failing list. □

The following corollary describes a case with certain symmetry where Lemma 2.1 can be applied. This special case is later used to indicate RS codes that contain failing lists.

COROLLARY 2.2. *Suppose there are* integers $\gamma > 0$ *and* $\gamma' \geq 0$ *that satisfy*

$$\binom{\ell+1}{r}\gamma + \binom{\ell+1}{r+1}\gamma' = n \qquad \textit{and} \qquad \binom{\ell-1}{r-2}\gamma + \binom{\ell-1}{r-1}\gamma' = n{-}d \tag{15}$$

(here $\tau_\ell(n,d)$ is an integer and its value is given by $\binom{\ell}{r}\gamma + \binom{\ell}{r+1}\gamma'$). Let $\mathcal{L} = \{\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_\ell\}$ be an (ℓ, r)-configuration with respect to a partition vector $(I_i)_i \| (I'_j)_j$ of $\langle n \rangle$, where $|I_i| = \gamma$ and $|I'_j| = \gamma'$. Let $\boldsymbol{v} \in F^n$ be the majority-vote word. Then $\mathcal{L}$ is a failing list in which $\mathsf{d}_H(\boldsymbol{c}_s, \boldsymbol{c}_t) = d$ for every $0 \leq s < t \leq \ell$ and $\mathsf{d}_H(\boldsymbol{c}_s, \boldsymbol{v}) = \tau_\ell(n,d)$ for every $0 \leq s \leq \ell$.

We point out that the failing lists described in [9], corresponding to cases where the relative minimum distance is $1 - \rho_r$, have a combinatorial structure which is a special case of the (ℓ, r)-configuration in Corollary 2.2, obtained when $\gamma' = 0$. As indicated in [9], the incidence structure $\boldsymbol{D}(\mathcal{L})$ in this case is a replication of the trivial (complete) balanced incomplete block design (BIBD) with parameters $(\ell{+}1, r, (n{-}d)/\gamma)$. In such a BIBD, the $n{=}\binom{\ell+1}{r}$ blocks correspond to all the distinct r-subsets of the point set $\mathcal{L}$, each pair of points appears in exactly $(n{-}d)/\gamma{=}\binom{\ell}{r}$ blocks, and each single point appears in exactly $(n{-}\tau_\ell(n,d))/\gamma{=}\binom{\ell}{r-1}$ blocks.

EXAMPLE 2.1. *Figure* 1 *presents a $(3,2)$-configuration of four words of length* 10 *over* GF(11) *and the respective majority-vote word $\boldsymbol{v}$. The words $\boldsymbol{c}_0, \boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3$ are codewords of a* $[10,4,7]$ *RS code whose code locators are* $0,5,6,4,2,1,7,3,9,8$. *The configuration forms a failing list since every two codewords agree on $n{-}d = 3$ positions and $\boldsymbol{v}$ agrees with every codeword on $\tau_3(10,7) = 4$ positions. Note that, for every two distinct $s,t \in \{0,1,2,3\}$, there is a unique position on which only $\boldsymbol{c}_s$ and $\boldsymbol{c}_t$ agree, and, for every three distinct $s,t,u \in \{0,1,2,3\}$, there is a unique position on which only $\boldsymbol{c}_s$, $\boldsymbol{c}_t$, and $\boldsymbol{c}_u$ agree. This list thus corresponds to the structure described in Corollary* 2.2, *where $\gamma = \gamma' = 1$.* □

$$\begin{array}{rcccccccccc}
\boldsymbol{c}_0 = & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\boldsymbol{c}_1 = & 0 & 0 & 2 & 3 & 0 & 4 & 4 & 5 & 10 & 1 \\
\boldsymbol{c}_2 = & 0 & 8 & 0 & 3 & 1 & 0 & 3 & 5 & 6 & 8 \\
\boldsymbol{c}_3 = & 6 & 0 & 0 & 3 & 8 & 5 & 0 & 1 & 10 & 8 \\
 & & & & & & & & & & \\
\boldsymbol{v} = & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 5 & 10 & 8
\end{array}$$

FIG. 1. $(3,2)$-*configuration over* GF(11).

EXAMPLE 2.2. *Figure* 2 *presents a $(4,3)$-configuration of five codewords of a* $[10,4,7]$ *RS code over* GF(16) *(the field is represented as polynomials over* GF(2) *modulo* $x^4 + x + 1$*, and the four polynomial coefficients of each element are written in hexadecimal notation). The rate,* $1 - d/n = 3/10$*, equals the boundary rate* $\rho_3(4)$*. This configuration follows the structure described in Corollary* 2.2*, where* $\gamma = 1$ *and* $\gamma' = 0$*, and it therefore forms a failing list. It can be easily verified that indeed every two codewords agree on* $n{-}d = 3$ *positions and* $\boldsymbol{v}$ *agrees with every codeword on* $\tau_4(10,7) = 4$ *positions. The list structure here corresponds to the (complete)* BIBD$(5,3,3)$ *(which has* 10 *blocks), and we show in what follows (Lemma* 4.4*) that, in fact,* every *failing list of five codewords in a* $(10, M, 7)$ *code over any alphabet* F *must have the form of a* BIBD$(5,3,3)$*. It follows that such a failing list cannot be realized over the binary alphabet, and by Propositions* 4.3 *and* 1.6*, it can be realized in RS codes over* GF(q) *if and only if* q *is a power of* 2 *not smaller than* 16*.* □

$$\begin{array}{rcccccccccc}
\boldsymbol{c}_0 = & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\boldsymbol{c}_1 = & 0 & 0 & 0 & 9 & \mathrm{f} & 2 & \mathrm{a} & 4 & \mathrm{f} & \mathrm{b} \\
\boldsymbol{c}_2 = & 0 & 9 & 1 & 0 & 0 & \mathrm{b} & \mathrm{a} & 4 & 7 & \mathrm{c} \\
\boldsymbol{c}_3 = & \mathrm{a} & 0 & 5 & 0 & \mathrm{a} & 0 & \mathrm{a} & \mathrm{a} & \mathrm{f} & \mathrm{c} \\
\boldsymbol{c}_4 = & \mathrm{c} & 4 & 0 & \mathrm{f} & 0 & 0 & 1 & 4 & \mathrm{f} & \mathrm{c}
\end{array}$$

FIG. 2. $(4,3)$-*configuration over* GF(16).

Fix a list size ℓ and a rational number $\delta \in (0,1]$. We claim that one can always extend ℓ to some admissible quadruple (ℓ, n, d, q) with $d/n = \delta$ such that $\ell{+}1$ words that form a failing list are contained in F^n (where F is an alphabet of size q). Indeed, replacing d by $n\delta$ in (15) (where r is uniquely determined by δ and ℓ) transforms (15) into a set of two homogeneous equations in the three unknowns γ, γ', and n. A nontrivial integer solution must then exist. For any value of q greater than $\ell{+}2{-}r$, we can find $\ell{+}1$ words in F^n that form an (ℓ, r)-configuration with respect to some partition vector $\mathcal{P} = (I_i)_i \| (I'_j)_j$ of $\langle n \rangle$, where $|I_i| = \gamma$ for $1 \le i \le \binom{\ell+1}{r}$, and $|I'_j| = \gamma'$ for $1 \le j \le \binom{\ell+1}{r+1}$. By Corollary 2.2, this is a failing list.

2.3. Necessary conditions on the existence of failing lists. Proposition 2.3 below motivates our interest in failing lists that form (ℓ, r)-configurations. It states that when $\tau_\ell(n,d)$ is an integer, namely, when (6) holds, every failing list of size $\ell{+}1$ is necessarily an (ℓ, r)-configuration. The sufficient condition for the existence of a failing list, as stated in Lemma 2.1, thus turns out to be necessary in cases where (6) holds.

PROPOSITION 2.3. *Let* ℓ*,* r*,* n*, and* d *be integers for which* (6) *holds, and let* $\mathcal{L}$ *be a failing list of size* $\ell{+}1$ *that is contained in an* (n, M, d) *code over* F*.*

N1. *The list* $\mathcal{L}$ *is an* (ℓ, r)*-configuration with respect to some partition vector* $\mathcal{P} = (I_i)_i \| (I'_j)_j$ *of* $\langle n \rangle$ *that satisfies conditions* (13)–(14) with equality.

N2. $\mathcal{P}$ *is proper (i.e., exactly* r *out of the* $\ell{+}1$ *words in* $\mathcal{L}$ *agree on every position) if and only if* $1 - d/n = r(r{-}1)/(\ell(\ell{+}1)) = \rho_r$*.*

N3. *If* $1 - d/n = r(r{-}1)/(\ell(\ell{+}1))$*, then* $q \ge \ell{+}1{-}r$*. Otherwise,* $q \ge \ell{+}2{-}r$*.*

Remark. Property N3 above is a restatement of Proposition 1.3.

Proof. Let $\mathcal{L} = \{\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_\ell\}$ be a failing list with the given parameters, and let $\boldsymbol{v}$ be the word in F^n for which $\mathsf{d}_H(\boldsymbol{c}_s, \boldsymbol{v}) \le \tau = \tau_\ell(n,d)$ for every $s \in \{0, 1, \ldots, \ell\}$. As in the proof of Theorem 1.2, we denote by x_μ, $\mu \in \langle n \rangle$, the number of words in $\mathcal{L}$ that agree with $\boldsymbol{v}$ on the μth position. By arguments similar to those in the proof of

Theorem 1.2, we get that

$$\sum_{\mu=1}^{n} x_\mu \ge (\ell{+}1)(n-\tau) \tag{16}$$

and

$$\sum_{\mu=1}^{n} \tbinom{x_\mu}{2} \le \tbinom{\ell+1}{2}(n{-}d)\ . \tag{17}$$

Let y be as in (11). Under the assumption that (6) holds, y must be an integer. When $1 - d/n = \rho_r$, we get $y = 0$; otherwise, $0 < y < n$. Regard $x_1, x_2, \ldots, x_n$ as integer variables that are constrained to satisfy (16) with equality. By [11, p. 526], the minimum of the sum $\sum_{\mu=1}^{n} \binom{x_\mu}{2}$ is attained when (and only when) y of the variables take the value $r{+}1$ while the rest take the value r; such an assignment satisfies (17) with equality. Since the minimum could only increase if we constrained the sum $\sum_{\mu=1}^{n} x_\mu$ to be larger, we have thus characterized the only feasible solutions to (16)–(17).

We now define the partition vector $\mathcal{P}$ that is stated in the lemma. For every subset S_i (respectively, S'_j) of $\{0, 1, \ldots, \ell\}$ of size r (respectively, $r{+}1$), let I_i (respectively, I'_j) be the set of positions on which the words in $\mathcal{L}_i = \{\boldsymbol{c}_s : s \in S_i\}$ (respectively, $\mathcal{L}_j = \{\boldsymbol{c}_s : s \in S'_j\}$)—and only these words—agree with $\boldsymbol{v}$.

Since the union of $\bigcup_i I_i$ and $\bigcup_j I'_j$ is necessarily $\langle n \rangle$, it follows that $\mathcal{P} = (I_i)_i \| (I'_j)_j$ is a partition vector. We have

$$\sum_{i\,:\,\{s,t\}\subseteq S_i} |I_i| \;+ \sum_{j\,:\,\{s,t\}\subseteq S'_j} |I'_j| \;\le\; n - \mathsf{d}_H(\boldsymbol{c}_s, \boldsymbol{c}_t)\ , \qquad 0 \le s < t \le \ell\ , \tag{18}$$

and

$$\sum_{i\,:\,s\in S_i} |I_i| \;+ \sum_{j\,:\,s\in S'_j} |I'_j| \;=\; n - \mathsf{d}_H(\boldsymbol{c}_s, \boldsymbol{v})\ , \qquad 0 \le s \le \ell\ . \tag{19}$$

Since $\mathcal{L}$ is a failing list, we can bound the right-hand side of (18) from above by $n{-}d$ and the right-hand side of (19) from below by $n - \tau$. This, in turn, implies that conditions (13)–(14) hold. Furthermore, since (16)–(17) hold with equality, we obtain

$$\tfrac{1}{2} \sum_{0\le s<t\le\ell} \Bigl(\sum_{i\,:\,\{s,t\}\subseteq S_i} |I_i| \;+ \sum_{j\,:\,\{s,t\}\subseteq S'_j} |I'_j| \Bigr) \;=\; \sum_{\mu=1}^{n} \tbinom{x_\mu}{2} = \tbinom{\ell+1}{2}(n{-}d)$$

and

$$\sum_{s=0}^{\ell} \Bigl(\sum_{i\,:\,s\in S_i} |I_i| \;+ \sum_{j\,:\,s\in S'_j} |I'_j| \Bigr) \;=\; \sum_{\mu=1}^{n} x_\mu \;=\; (\ell{+}1)(n-\tau)\ .$$

It follows that conditions (13)–(14) hold with equality, and so does (18). The equality in (18) implies that when $x_\mu = r$ (respectively, $x_\mu = r{+}1$), there are exactly $\binom{r}{2}$ (respectively, $\binom{r+1}{2}$) different pairs of words in $\mathcal{L}$ that agree on their μth coordinate. In particular, a word in $\mathcal{L} \setminus \mathcal{L}_i$ (respectively, $\mathcal{L} \setminus \mathcal{L}'_j$) does not agree on any position in I_i (respectively, I'_j) with any other word in $\mathcal{L}$. We conclude that $\mathcal{L}$ is an (ℓ, r)-configuration with respect to the partition vector $\mathcal{P}$, and property N1 is thus proved. Recalling that $y{=}0$ when $1 - d/n = \rho_r$, property N2 is proved as well. Property N3 is implied by properties N1–N2 and by the definition of an (ℓ, r)-configuration. □

2.4. Constant-weight codes, block designs, and failing lists. One necessary condition on the existence of failing lists is given in Proposition 2.4 below by means of constant-weight codes. If F is an additive group, then an (n, d, w) constant-weight code over F is a subset of F^n such that the Hamming weight (i.e., the number of nonzero components) of every codeword is w and the minimum Hamming distance between different codewords is d (see also [11, p. 524]).

PROPOSITION 2.4. *Let (ℓ, n, d, q) be an admissible quadruple, let r be the unique integer such that $1-d/n \in [\rho_r, \rho_{r+1})$, and assume that* (6) *holds. Suppose that a failing list is contained in some (n, M, d) code over an additive group F of size q. Then a (possibly different) failing list forms an $(n, d, \tau_\ell(n, d))$ constant-weight code $\bar{\mathcal{C}}$ over F, consisting of $\ell+1$ codewords. The Hamming distance between different codewords in $\bar{\mathcal{C}}$ is* exactly *d.*

Proof. Let $\mathcal{L} = \{\boldsymbol{c}_s\}_{s=0}^{\ell}$ be the failing list, and let $\boldsymbol{v}$ be as in the proof of Proposition 2.3. By property N1 of that proposition, the set $\{\boldsymbol{c}_0-\boldsymbol{v}, \boldsymbol{c}_1-\boldsymbol{v}, \ldots, \boldsymbol{c}_\ell-\boldsymbol{v}\}$ forms the required constant-weight code over F. □

Let $\mathcal{L}$ be a failing list as in Proposition 2.3. We consider the incidence structure $\boldsymbol{D}(\mathcal{L}) = (\mathcal{L}, \mathcal{B}, \mathcal{M})$ as a generalization of a BIBD$(\ell+1, r, n-d)$, referred to as a *quasi*-BIBD and denoted QBIBD$(\ell+1, r, n-d; n)$. For an introduction on BIBDs, see [2], [8, Ch. 10], and [11, sect. 2.5]. In a QBIBD, similarly to a BIBD, every pair of points appears in exactly $n-d$ blocks (the incidence structure is pairwise balanced), and each single point appears in exactly $n-\tau_\ell(n, d)$ blocks. However, in a QBIBD, y blocks are of size $r + 1$, where y is defined by (11), and the remaining $n-y > 0$ blocks are of size r. In addition, repeated blocks are allowed in a QBIBD.

Note that the number of blocks n appears as a parameter in the definition of a QBIBD since it is not uniquely determined by the other three parameters. However, the following connection between the parameters must hold:

$$\binom{r}{2}n \le \binom{\ell+1}{2}(n-d) < \binom{r+1}{2}n \,. \tag{20}$$

When the left inequality in (20) holds with equality (i.e., the code relative minimum distance is $1 - \rho_r$), the n blocks are all of size r.

Some useful properties of a BIBD, such as Fisher's inequality (see, for example, [2, p. 81]), also hold for a QBIBD, as stated in the following lemma. The proof is essentially the same as in the case of a BIBD, and it is included for the sake of completeness.

LEMMA 2.5. *In a* QBIBD$(\ell+1, r, n-d; n)$*, there are at least* $\ell+1$ distinct *blocks. In particular, $\ell+1 \le n$.*

Proof. Let $\boldsymbol{D}(\mathcal{L}) = (\mathcal{L}, \mathcal{B}, \mathcal{M})$ be an incidence structure of a QBIBD$(\ell+1, r, n-d; n)$. The entries of the $(\ell+1) \times (\ell+1)$ matrix $\mathcal{M}\mathcal{M}^T$ are given by

$$(\mathcal{M}\mathcal{M}^T)_{s,t} = \begin{cases} n-\tau_\ell(n,d) & \text{if } s = t, \\ n-d & \text{if } s \ne t, \end{cases}$$

and, so,

$$\det \mathcal{M}\mathcal{M}^T = (d-\tau_\ell(n,d))^\ell \cdot (\ell(n-d)+n - \tau_\ell(n,d)) \ne 0 \,.$$

It follows that $\mathcal{M}\mathcal{M}^T$ contains at least $\ell+1$ linearly independent—and hence distinct—columns. □

The following corollary is implied by Proposition 2.3 and Lemma 2.5.

COROLLARY 2.6. *Let ℓ, r, n, and d be integers for which* (6) *holds. Then a failing list of size $\ell+1$ is contained in an (n, M, d) code over some alphabet F only* *if there exists a* QBIBD$(\ell{+}1, r, n{-}d; n)$*. In particular,* $\ell{+}1 \le n$ *whenever a failing list* $\mathcal{L}$ *exists.*

The next lemma deals with the special case $n = \ell{+}1$.

Lemma 2.7. *A* QBIBD$(n, r, n{-}d; n)$ *is a (symmetric)* BIBD$(n, r, n{-}d)$*.*

Proof. By Lemma 2.5, the n blocks are all distinct. Now, in a QBIBD$(n, r, n{-}d; n)$, each point appears in $n{-}\tau_\ell(n, d)$ blocks and, so, $\tau_\ell(n, d)$ is an integer. The divisibility condition (6), which necessarily holds here, becomes

$$2r \quad \text{divides} \quad r(r{+}1) + (n{-}1)(n{-}d)\,.$$

By (20), we also require that

$$r(r{-}1) \le (n{-}1)(n{-}d) < r(r{+}1)\,.$$

The above two constraints are satisfied only if $(n{-}1)(n{-}d) = r(r{-}1)$, implying that the n *distinct* blocks are all of the *same size* r. The QBIBD is thus a BIBD. □

Proposition 2.8 below deals with list sizes $\ell \ge n{-}1$. In particular, it states that when $\ell = n{-}1$, the GS bound can be attained only when there is a symmetric BIBD with parameters $(n, r, n{-}d)$. Such a design consists of n "points" and n "blocks" of size r, where each pair of distinct points appears in exactly $n{-}d$ blocks.

Proposition 2.8. *Let* ℓ, r, n, d, q *be as in Proposition* 2.4*.*

B1. *If* $n = \ell{+}1$, *then* $\Delta_\ell(n, d; q) = \tau_\ell(n, d){-}1$ only *when there is a* BIBD$(n, r, n{-}d)$ *with* $r(r{-}1) = (n{-}1)(n{-}d)$*.*

B2. *If* $n < \ell{+}1$, *then* $\Delta_\ell(n, d; q) \ge \tau_\ell(n, d)$*.*

Proof. Combine Corollary 2.6 and Lemma 2.7. □

Necessary conditions on the parameters of a symmetric BIBD were given by Bruck, Chowla, and Ryser (see [2, p. 100] or [8, p. 133]). It follows from Proposition 2.8 that whenever these conditions are not satisfied by $(n, r, n{-}d)$, no (n, M, d) code attains the GS bound with equality. For example, since there is no BIBD$(22, 7, 2)$, we obtain for every alphabet size q

$$\Delta_{21}(22, 20; q) \ge 15 = \tau_{21}(22, 20)\ .$$

Similarly,

$$\Delta_{42}(43, 42; q) \ge 36 = \tau_{42}(43, 42)\,.$$

3. Realizing failing lists in RS codes. Throughout this section we fix the finite field $F = \mathrm{GF}(q)$, the list size ℓ, and an $[n, k, d]$ RS code $\mathcal{C}$ over F with a set of code locators $\{\alpha_1, \alpha_2, \ldots, \alpha_n\} \subseteq F$. We let r be the unique integer such that $1 - \delta = (k{-}1)/n \in [\rho_r, \rho_{r+1})$.

Suppose that $\mathcal{C}$ contains a set $\mathcal{L} = \{\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_\ell\}$ which is an (ℓ, r)-configuration with respect to some partition vector $\mathcal{P}$ for which (13)–(14) are satisfied. Without loss of generality, assume that $\boldsymbol{c}_0$ is the zero codeword (otherwise, subtract $\boldsymbol{c}_0$ from each $\boldsymbol{c}_s$ to obtain another (ℓ, r)-configuration with respect to $\mathcal{P}$). For every two indexes s, t such that $0 \le s < t \le \ell$, the difference $\boldsymbol{c}_s{-}\boldsymbol{c}_t$ is a codeword that is obtained by evaluating a polynomial of degree $\le k{-}1$ at the code locators. We denote this polynomial by $a_{s,t} \cdot f_{s,t}(x)$, where $a_{s,t} \in F \setminus \{0\}$ and $f_{s,t}(x)$ is a monic polynomial of degree $\le k{-}1$.

For every subset $S_i \subseteq \{0, 1, \ldots, \ell\}$ of size r and every subset S'_j of size $r{+}1$, define

$$A_{S_i}(x) = \prod_{\mu \in I_i} (x - \alpha_\mu) \qquad \text{and} \qquad A_{S'_j}(x) = \prod_{\mu \in I'_j} (x - \alpha_\mu)\ . \tag{21}$$

By the definition of an (ℓ, r)-configuration with respect to a partition vector, it follows that the polynomials $f_{s,t}(x)$ are given by

$$f_{s,t}(x) = \prod_{i\,:\,\{s,t\}\subseteq S_i} A_{S_i}(x) \cdot \prod_{j\,:\,\{s,t\}\subseteq S'_j} A_{S'_j}(x) \tag{22}$$

(a product over an empty set is defined as 1). A necessary condition on the existence of the configuration $\mathcal{L}$ in $\mathcal{C}$ is

$$a_{s,t} \cdot f_{s,t}(x) = a_{0,s} \cdot f_{0,s}(x) - a_{0,t} \cdot f_{0,t}(x) \ , \quad 0 < s < t \le \ell \ . \tag{23}$$

Conversely, suppose that $\mathcal{P} = (I_i)_i \| (I'_j)_j$ is a partition vector that satisfies (13)–(14), and let the polynomials $f_{s,t}(x)$ be defined by (22). If there are nonzero constants $a_{s,t} \in F$ that satisfy (23), then an (ℓ, r)-configuration with respect to $\mathcal{P}$ exists. Based on Lemma 2.1 and Proposition 2.3, the following lemma is obtained.

Lemma 3.1. *Let $(I_i)_i \| (I'_j)_j$ be a partition vector of $\langle n \rangle$ that satisfies* (13)–(14) *and let the polynomials $f_{s,t}(x)$, $0 \le s < t \le \ell$, be defined by* (22).

(a) *If there are $\binom{\ell+1}{2}$ nonzero constants $a_{s,t} \in F$ such that* (23) *holds, then $\mathcal{C}$ contains a failing list that consists of the zero codeword and the ℓ codewords*

$$(f(\alpha_1)\ f(\alpha_2)\ \cdots\ f(\alpha_n)) \ , \quad f(x) \in \{a_{0,s} \cdot f_{0,s}(x)\}_{s=1}^{\ell} \ .$$

(b) *In cases where* (6) *holds, the sufficient conditions in part* (a) *for the existence of a failing list of size $\ell{+}1$ are also necessary, and each polynomial $f_{s,t}(x)$ has degree $k{-}1$.*

3.1. The difference condition and simple sets of polynomials. Three monic polynomials $f(x) = \sum_i f_i x^i$, $g(x) = \sum_i g_i x^i$, and $h(x) = \sum_i h_i x^i$ are said to satisfy the *difference condition* if there are $\tilde{f}, \tilde{g}, \tilde{h} \in F \setminus \{0\}$ for which

$$\tilde{h} \cdot h(x) = \tilde{f} \cdot f(x) - \tilde{g} \cdot g(x) \ . \tag{24}$$

Observe that every three polynomials in (23) must satisfy the difference conditions.

Lemma 3.2. *Three distinct monic polynomials of the same degree e, $f(x) = \sum_{i \le e} f_i x^i$, $g(x) = \sum_{i \le e} g_i x^i$, and $h(x) = \sum_{i \le e} h_i x^i$, satisfy the difference condition if and only if*

$$(f_i - h_i)(f_j - g_j) = (f_i - g_i)(f_j - h_j) \quad \textit{for every} \quad 0 \le i, j \le e \ . \tag{25}$$

Proof. The difference condition is satisfied if and only if there are $\tilde{f}, \tilde{g}, \tilde{h} \in F \setminus \{0\}$ for which

$$\tilde{h} \cdot h_i = \tilde{f} \cdot f_i - \tilde{g} \cdot g_i, \quad 0 \le i \le e \ , \tag{26}$$

and since $f(x)$, $g(x)$, and $h(x)$ are monic polynomials of the same degree e, we obtain, in particular, that

$$\tilde{h} = \tilde{f} - \tilde{g} \ . \tag{27}$$

A nontrivial solution for $\tilde{f}, \tilde{g}, \tilde{h}$ exists if and only if the following matrix is singular for every $0 \le i \le j \le e$:

$$\begin{pmatrix} 1 & -1 & -1 \\ f_i & -g_i & -h_i \\ f_j & -g_j & -h_j \end{pmatrix} .$$

This matrix, in turn, is singular if and only if (25) holds.

Now, the values of $\tilde{f}, \tilde{g}, \tilde{h}$ must be all nonzero in every nontrivial solution of (24), as required by the difference condition; by (27), it is impossible that exactly two of them are zero, and if only one is zero, then, by combining (26) and (27), we obtain that two out of three polynomials $f(x)$, $g(x)$, and $h(x)$ are identical; this, however, contradicts our assumption that these polynomials are distinct. □

Our constructions that realize (22)–(23) will have a special structure defined next. A set of polynomials of degree e over F is *simple over a set* $U \subseteq F$ if the following three conditions hold:

S1. Each polynomial has e simple roots in U.

S2. Every two distinct polynomials in the set are relatively prime.

S3. The polynomials differ only in the ith coefficient for some i. For example, they differ only in their constant term.

Corollary 3.3. *Every three polynomials in a simple set satisfy the difference condition.*

Proof. Let $f(x)$, $g(x)$, and $h(x)$ be three polynomials of degree e in a simple set. By property S3 of simple sets, $(f_i - h_i)(f_j - g_j) = (f_i - g_i)(f_j - h_j) = 0$ for every distinct i, j such that $0 \leq i, j \leq e$. Obviously, for $i = j$ we have $(f_i - h_i)(f_j - g_j) = (f_i - g_i)(f_j - h_j)$. By Lemma 3.2, the difference condition is satisfied. □

3.2. Rates above $2/(\ell(\ell{+}1))$. Lemma 3.4 below provides a *sufficient* condition on the existence of failing lists of size $\ell{+}1$ in $\mathcal{C}$. The statement of the lemma makes use of the following notation. Let $\mathcal{P} = (I_i)_i \| (I'_j)_j$ be a partition vector of $\langle n \rangle$ and let $f_{s,t}(x)$, $0 \leq s < t \leq \ell$, be defined by (22). For every s, t, u such that $0 \leq s < t < u \leq \ell$, define

$$(28) \qquad \begin{aligned} g_{s,t,u}(x) &= \gcd(f_{s,t}(x), f_{s,u}(x), f_{t,u}(x)) \\ &= \prod_{i\,:\,\{s,t,u\} \subseteq S_i} A_{S_i}(x) \cdot \prod_{j\,:\,\{s,t,u\} \subseteq S'_j} A_{S'_j}(x). \end{aligned}$$

Lemma 3.4. *Let $\mathcal{P} = (I_i)_i \| (I'_j)_j$ be a partition vector of $\langle n \rangle$ for which* (13)–(14) *hold, and let $f_{s,t}(x)$ and $g_{s,t,u}(x)$ be the polynomials defined by* (22) *and* (28), *respectively. A failing list of size $\ell{+}1$ is contained in $\mathcal{C}$ if the following two conditions hold:*

- *For every $0 \leq s < t \leq \ell$, the polynomials $f_{0,s}(x)$, $f_{0,t}(x)$, and $f_{s,t}(x)$ satisfy the difference condition, and*
- *$g_{s,t,u}(x)$ does not divide $f_{0,s}(x)$ for every $0 < s < t < u \leq \ell$.*

Proof. We show that the sufficient conditions of Lemma 3.1(a) hold. If $f_{0,s}(x)$, $f_{0,t}(x)$, and $f_{s,t}(x)$ satisfy the difference condition, then, by definition, there must be nonzero $a_{0,s}, a_{0,t}, a_{s,t} \in F$ such that

$$(29) \qquad a_{s,t} \cdot f_{s,t}(x) = a_{0,s} \cdot f_{0,s}(x) \, - \, a_{0,t} \cdot f_{0,t}(x) \, .$$

For the case where $\ell = 2$, we are done.

Turning to larger values of ℓ, we need to show that the same coefficient $a_{0,s}$ multiplies $f_{0,s}(x)$ in (23), independently of t. Given $s \in \langle \ell{-}2 \rangle$, consider any indexes t and u such that $s < t < u \leq \ell$. There must be nonzero $a_{0,s}, a_{0,t}, a_{s,t} \in F$ satisfying (29) and, by the same arguments, there must be nonzero $a_{0,u}, a_{t,u}, a_{s,u}, a'_{0,s} \in F$ such that

$$(30) \qquad a_{t,u} \cdot f_{t,u}(x) = a_{0,t} \cdot f_{0,t}(x) \, - \, a_{0,u} \cdot f_{0,u}(x) \text{ and}$$

$$(31) \qquad a_{s,u} \cdot f_{s,u}(x) = a'_{0,s} \cdot f_{0,s}(x) \, - \, a_{0,u} \cdot f_{0,u}(x) \, .$$

Subtracting (31) from the sum of (29) and (30) results in

$$a_{s,t} \cdot f_{s,t}(x) + a_{t,u} \cdot f_{t,u}(x) - a_{s,u} \cdot f_{s,u}(x) = (a_{0,s} - a'_{0,s}) \cdot f_{0,s}(x)\ . \tag{32}$$

Clearly, $g_{s,t,u}(x)$ divides the left-hand side of (32). However, according to the assumptions of the lemma, $g_{s,t,u}(x)$ does not divide $f_{0,s}(x)$. We therefore conclude that $a'_{0,s} = a_{0,s}$, i.e., the same coefficient $a_{0,s}$ does indeed multiply $f_{0,s}(x)$ in (23), independently of t. □

Corollaries 3.5 and 3.6 below are derived from Lemma 3.4 and are used in section 4 to indicate families of RS codes that contain failing lists of size $\ell{+}1$. Corollary 3.5 covers only the high-rate range $(k{-}1)/n \geq 1 - (2/(\ell{+}1))\ (= \rho_\ell)$, while Corollary 3.6 applies to $(k{-}1)/n > 2/(\ell(\ell{+}1))\ (= \rho_2)$.

COROLLARY 3.5. *Let the positive integer triple* (ℓ, n, k) *be such that* $(k{-}1)/n \geq 1 - (2/(\ell{+}1))$. *A failing list is contained in* $\mathcal{C}$ *if there is some partition vector* $\mathcal{P} = (I_1, I_2, \ldots, I_{\ell+1}, I'_1)$ *of* $\langle n \rangle$ *such that the following hold:*

(a) (13)–(14) *are satisfied, with equality in* (13), *and*

(b) *for every* $0 \leq s < t \leq \ell$ *in* (22), *the respective polynomials* $f_{0,s}(x)$, $f_{0,t}(x)$, *and* $f_{s,t}(x)$, *each of degree* $k{-}1$, *are* distinct *and satisfy the difference condition.*

Proof. We show that whenever $r = \ell > 2$, for every $0 < s < t < u \leq \ell$ the polynomial $g_{s,t,u}(x)$ in (28) does not divide $f_{0,s}(x)$. The existence of a failing list will then follow from Lemma 3.4.

Without loss of generality we can assume that the sets S_i are defined so that $S_1 = \{1, 2, \ldots, \ell\}$. For every $0 < s < t < u$, the polynomial $g_{s,t,u}(x)$ does not divide $f_{0,s}(x)$ if and only if $\deg A_{S_1}(x) > 0$. Assume to the contrary that $\deg A_{S_1}(x) = 0$. Since $A_{S_1}(x) = f_{s,t}(x)/g_{0,s,t}(x)$, it then follows that $\deg g_{0,s,t}(x) = k{-}1$; therefore, $f_{0,s}(x) = g_{0,s,t}(x) = f_{0,t}(x)$, contradicting our assumption that $f_{0,s}(x)$ and $f_{0,t}(x)$ are distinct. □

COROLLARY 3.6. *Let the positive integer triple* (ℓ, n, k) *be such that* $(k{-}1)/n > 2/(\ell(\ell{+}1))$, *and let* $r > 1$, $\gamma > 0$, *and* $\gamma' \geq 0$ *be integers for which* (15) *holds. Let* $\mathcal{P} = (I_i)_i \| (I'_j)_j$ *be a partition vector of* $\langle n \rangle$ *in which* $|I_i| = \gamma$ *and* $|I'_j| = \gamma'$, *and let* $f_{s,t}(x)$ *and* $g_{s,t,u}(x)$ *be the polynomials defined by* (22) *and* (28), *respectively. Suppose that for every* $0 \leq s < t \leq \ell$, *there is a polynomial divisor* $\lambda_{s,t}(x)$ *of* $g_{0,s,t}(x)$ *for which the set* $\{f_{0,s}(x)/\lambda_{s,t}(x), f_{0,t}(x)/\lambda_{s,t}(x), f_{s,t}(x)/\lambda_{s,t}(x)\}$ *is simple over the set of code locators of* $\mathcal{C}$. *Then* $\mathcal{C}$ *contains a failing list of size* $\ell{+}1$.

Proof. By Lemma 3.4, it suffices to show that $g_{s,t,u}(x)$ does not divide $f_{0,s}(x)$ for every $0 < s < t < u \leq \ell$. If $2 < r \leq \ell$, there is a nonempty partition element I_i in $\mathcal{P}$ that corresponds to a subset $S_i \subseteq \{0, 1, \ldots, \ell\}$ such that $\{s, t, u\} \subseteq S_i$ while $0 \notin S_i$. In this case, $A_{S_i}(x)$ divides $g_{s,t,u}(x)$, but it does not divide $f_{0,s}(x)$; therefore, $g_{s,t,u}(x)$ does not divide $f_{0,s}(x)$, as required.

Assume now that $2 = r < \ell$. Since $(k{-}1)/n > 2/(\ell(\ell{+}1))$, there must exist a nonempty partition element I'_j in $\mathcal{P}$ that corresponds to a subset $S'_j = \{s, t, u\}$ of $\{0, 1, \ldots, \ell\}$. Since $|I'_j| = \gamma' > 0$, the polynomial $A_{S'_j}(x)$ divides $g_{s,t,u}(x)$ but not $f_{0,s}(x)$; thus, $g_{s,t,u}(x)$ does not divide $f_{0,s}(x)$, as required. □

3.3. The low-rate range: $0 < (k{-}1)/n \leq 2/(\ell(\ell{+}1))$. Suppose that the rate of $\mathcal{C}$ satisfies $(k{-}1)/n < \rho_2 = 2/(\ell(\ell{+}1))$ and that $\mathcal{C}$ contains an $(\ell, 1)$-configuration $\mathcal{L}$. At most two out of the $\ell{+}1$ words in $\mathcal{L}$ agree on every position and, so, (22) becomes $f_{s,t}(x) = f_{s,t}(x)/g_{0,s,t}(x) = A_{S'_j}(x)$ for $S'_j = \{s, t\}$. Suppose now that $(k{-}1)/n = \rho_2$ and that $\mathcal{L}$ is an $(\ell, 2)$-configuration; here, $f_{s,t}(x) = f_{s,t}(x)/g_{0,s,t}(x) = A_{S_i}(x)$ for $S_i = \{s, t\}$. In both cases, the set $\{f_{s,t}(x)\}_{s,t}$ already satisfies conditions S1 and S2 for being simple over the set of code locators of $\mathcal{C}$.

However, it turns out that when $\ell > 2$ in any of those two cases, taking the set $\{f_{s,t}(x)\}_{s,t}$ to be simple over the set of code locators does not guarantee the existence of multipliers $\{a_{s,t}\}_{s,t}$ for which (23) holds. An auxiliary condition on the coefficients of $\{f_{s,t}(x)\}_{s,t}$ is needed in this case, as stated in the following lemma.

LEMMA 3.7. *Let the positive integer triple (ℓ, n, k) be such that $(k-1)/n \le 2/(\ell(\ell+1))$. Suppose there exists a set $\{f^*_{0,1}(x), f^*_{0,2}(x), \ldots, f^*_{\ell-1,\ell}(x)\}$ of $\binom{\ell+1}{2}$ distinct polynomials of degree $k-1$ over F that is simple over a subset U of size $\binom{\ell+1}{2}(k-1)$ of the set of code locators of $\mathcal{C}$. Let e be the (unique) coefficient index in which the polynomials differ, and denote by $\psi_{s,t}$ the eth coefficient of $f^*_{s,t}(x)$. Assume that when $\ell > 2$, the coefficients $\psi_{s,t}$ satisfy the $\binom{\ell-1}{2}$ equations*

$$\begin{aligned}(33)\qquad &(\psi_{1,s}-\psi_{0,1})(\psi_{1,t}-\psi_{0,t})(\psi_{s,t}-\psi_{0,s})\\ &\quad = (\psi_{1,s}-\psi_{0,s})(\psi_{1,t}-\psi_{0,1})(\psi_{s,t}-\psi_{0,t}) \ , \quad 1 < s < t \le \ell.\end{aligned}$$

Then there exist nonzero $a_{0,1}, a_{0,2}, \ldots, a_{0,\ell} \in F$ such that the zero word and the ℓ words

$$(f(\alpha_1)\ f(\alpha_2)\ \cdots\ f(\alpha_n)) \ , \quad f(x) \in \{a_{0,s} \cdot f^*_{0,s}(x)\}_{s=1}^{\ell}$$

form a failing list in $\mathcal{C}$.

Proof. Our proof is based on Lemma 3.1(a). To this end, we first find a partition vector $\mathcal{P} = (I_i)_i \| (I'_j)_j$ of $\langle n \rangle$ that satisfies (13)–(14), and that allows us to express the polynomials $f^*_{s,t}(x)$ in the form (22). When $(k-1)/n = \rho_2$, we select $\mathcal{P}$ to be proper, and for every $S_i = \{s,t\}$ we let $I_i = \{\mu : f^*_{s,t}(\alpha_\mu) = 0\}$.

When $(k-1)/n < \rho_2$, we select $\mathcal{P} = (I_i)_i \| (I'_j)_j$ so that for $S'_j = \{s,t\}$ the partition element I'_j is given by $\{\mu : f^*_{s,t}(\alpha_\mu) = 0\}$. Each of the $\ell+1$ partition elements I_i, which correspond to singleton subsets S_i, contains at least $\lfloor (n-|U|)/(\ell+1) \rfloor$ of the remaining elements of $\langle n \rangle$. Since the various polynomials $f^*_{s,t}(x)$ are all distinct, $\mathcal{P}$ is indeed a partition vector. It is also clear that $\mathcal{P}$ satisfies (13) with equality. As for (14), for every $s = 0, 1, \ldots, \ell$,

$$\sum_{i:s \in S_i} |I_i| + \sum_{j:s \in S'_j} |I'_j| \ge \ell(k-1) + \left\lfloor \tfrac{n-|U|}{\ell+1} \right\rfloor = n - \left\lceil \tfrac{\ell n}{\ell+1} - \tfrac{\ell(k-1)}{2} \right\rceil = n - \lceil \tau_\ell(n,k) \rceil \ .$$

Given the partition vector $\mathcal{P}$, we have $f^*_{s,t}(x) = f_{s,t}(x)$, where $f_{s,t}(x)$ are given by (22). By Lemma 3.1(a), all we still need to show is that there are nonzero coefficients $a_{s,t}$, $0 \le s < t \le \ell$, for which (23) holds. We distinguish between three cases, according to the value of ℓ (omitting the obvious case $\ell = 1$).

Case 1 ($\ell = 2$). The three polynomials $f_{0,1}(x)$, $f_{0,2}(x)$, and $f_{1,2}(x)$ satisfy condition S3 of a simple set; therefore, by Lemma 3.2, they satisfy the difference condition.

Case 2 ($\ell = 3$). Since $f_{0,1}(x), f_{0,2}(x), \ldots, f_{2,3}(x)$ satisfy condition S3, the set of linear equations (23) has a nontrivial solution for the unknowns $a_{0,1}, a_{0,2}, \ldots, a_{2,3}$ if and only if there is a nontrivial solution for the following set of equations:

$$(34)\qquad \begin{pmatrix} 1 & -1 & 0 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 & -1 \\ \psi_{0,1} & -\psi_{0,2} & 0 & -\psi_{1,2} & 0 & 0 \\ \psi_{0,1} & 0 & -\psi_{0,3} & 0 & -\psi_{1,3} & 0 \\ 0 & \psi_{0,2} & -\psi_{0,3} & 0 & 0 & -\psi_{2,3} \end{pmatrix} \begin{pmatrix} a_{0,1} \\ a_{0,2} \\ a_{0,3} \\ a_{1,2} \\ a_{1,3} \\ a_{2,3} \end{pmatrix} = \mathbf{0} \ .$$

However, the determinant of the matrix in (34) is zero if and only if

$$(\psi_{1,2}-\psi_{0,1})(\psi_{1,3}-\psi_{0,3})(\psi_{2,3}-\psi_{0,2}) \;=\; (\psi_{1,2}-\psi_{0,2})(\psi_{1,3}-\psi_{0,1})(\psi_{2,3}-\psi_{0,3})\;;$$

this is condition (33) for $\ell = 3$. Furthermore, if one of the elements $a_{0,1}, a_{0,2}, \ldots, a_{2,3}$ is zero, then either all these elements are zero, or else $\psi_{s',t'} = \psi_{s'',t''}$ for some $(s',t') \neq (s'',t'')$, where $0 \le s' < t' \le 3$ and $0 \le s'' < t'' \le 3$; yet, the latter contradicts our assumption that $f_{0,1}(x), f_{0,2}(x), \ldots, f_{2,3}(x)$ are all distinct. Therefore, in a nontrivial solution for $a_{0,1}, a_{0,2}, \ldots, a_{2,3}$, all these elements are nonzero.

Case 3 ($\ell > 3$). Fix some s in the range $1 < s \le \ell{-}2$, and consider another index t in the range $s < t \le \ell{-}1$. Following the analysis of Case 2, there must exist nonzero $a_{0,1}, a_{0,s}, a_{0,t}, a_{1,s}, a_{1,t}, a_{s,t} \in F$ such that

$$\begin{aligned} a_{s,t}\cdot f_{s,t}(x) &= a_{0,s}\cdot f_{0,s}(x) - a_{0,t}\cdot f_{0,t}(x)\;,\\ a_{1,s}\cdot f_{1,s}(x) &= a_{0,1}\cdot f_{0,1}(x) - a_{0,s}\cdot f_{0,s}(x)\;,\\ a_{1,t}\cdot f_{1,t}(x) &= a_{0,1}\cdot f_{0,1}(x) - a_{0,t}\cdot f_{0,t}(x)\;. \end{aligned} \tag{35}$$

Let u be in the range $t < u \le \ell$; there are five nonzero coefficients $a'_{0,s}, a'_{1,s}, a_{0,u}, a_{1,u}, a_{s,u}$ such that

$$\begin{aligned} a_{s,u}\cdot f_{s,u}(x) &= a'_{0,s}\cdot f_{0,s}(x) - a_{0,u}\cdot f_{0,u}(x)\;,\\ a'_{1,s}\cdot f_{1,s}(x) &= a_{0,1}\cdot f_{0,1}(x) - a'_{0,s}\cdot f_{0,s}(x)\;,\\ a_{1,u}\cdot f_{1,u}(x) &= a_{0,1}\cdot f_{0,1}(x) - a_{0,u}\cdot f_{0,u}(x)\;. \end{aligned} \tag{36}$$

Combining (35) and (36) results in

$$(a_{1,s} - a'_{1,s})\cdot f_{1,s}(x) \;=\; (a'_{0,s} - a_{0,s})\cdot f_{0,s}(x)\;,$$

and since $f_{1,s}(x)$ and $f_{0,s}(x)$ are relatively prime, it follows that $a'_{0,s} = a_{0,s}$ and $a'_{1,s} = a_{1,s}$. Hence, the same nonzero constant $a_{0,s}$ multiplies $f_{0,s}(x)$ in (23), independently of t. □

4. Proof of main results for RS codes.

In this section (starting from subsection 4.3), we prove Propositions 1.4–1.7. We use the tools developed in section 3 and additional tools presented in the following two subsections.

4.1. Properties of $\Delta_\ell^{\mathrm{RS}}(n, d\,; q)$.

Proposition 4.1 below describes some simple relations satisfied by $\Delta_\ell^{\mathrm{RS}}(n, d; q)$.

PROPOSITION 4.1. *Let (ℓ, n, d, q) be an RS-admissible quadruple. Then*
(a) $\Delta_\ell^{RS}(n{-}1, d{-}1; q) \le \Delta_\ell^{RS}(n, d; q) \le \Delta_\ell^{RS}(n{-}1, d{-}1; q) + 1$ *for* $d > 1$, *and*
(b) $\Delta_\ell^{RS}(n, d; q) \le \Delta_\ell^{RS}(n{-}1, d; q)$ *for* $d < n$.

Proof. Part (a): Let $\mathcal{C}$ be an $[n,k,d]$ RS code over GF(q), where $k < n$ ($d > 1$), and let $\mathcal{C}'$ be obtained by deleting the last coordinate from each codeword of $\mathcal{C}$. A list-ℓ decoder for $\mathcal{C}$ can be obtained by truncating the last coordinate from the received word and applying a list-ℓ decoder for $\mathcal{C}'$ to the resulting word. Hence, $\Delta_\ell(\mathcal{C}) \ge \Delta_\ell(\mathcal{C}')$, and, so, $\Delta_\ell(n, d; q) \ge \Delta_\ell(n{-}1, d{-}1; q)$. On the other hand, a list-ℓ decoder for $\mathcal{C}'$ can be obtained by appending an arbitrary nth coordinate to the received word, followed by an application of a list-ℓ decoder for $\mathcal{C}$. Therefore, $\Delta_\ell(\mathcal{C}') \ge \Delta_\ell(\mathcal{C}) - 1$ and, since $\mathcal{C}'$ can be any $[n{-}1, k, d{-}1]$ RS code, $\Delta_\ell(n{-}1, d{-}1; q) \ge \Delta_\ell(n, d; q) - 1$.

Part (b): Every $[n{-}1, k{-}1, d]$ RS code $\mathcal{C}$ over GF(q) with $n \le q$ can be extended to an $[n, k, d]$ (generalized) RS code $\overline{\mathcal{C}}$ over GF(q) by adding one column to the parity-check matrix of $\mathcal{C}$; (see [11, sect. 10.8]). Therefore, a list-ℓ decoder for $\mathcal{C}$ can be obtained by appending a zero coordinate to the received word and then applying a list-ℓ decoder for $\overline{\mathcal{C}}$. Hence, $\Delta_\ell(\mathcal{C}) \ge \Delta_\ell(\overline{\mathcal{C}})$ and, so, $\Delta_\ell(n{-}1, d; q) \ge \Delta_\ell(n, d; q)$. □

4.2. Types of simple sets of polynomials. In some of our proofs, we will use two types of sets of polynomials that are simple over certain sets U, as follows.

Type 1. Assume that $e \,|\, q{-}1$ and let α be a primitive element in $F = \mathrm{GF}(q)$. The set

$$\Big\{ x^e - \alpha^{ei} \;:\; 0 \le i < (q{-}1)/e \Big\}$$

is simple over $F \setminus \{0\}$.

Type 2. Assume that $q = p^h$ and $e = p^b < q$. If we regard $F = \mathrm{GF}(q)$ as a linear space over $\mathrm{GF}(p)$, then the p^b elements of every b-dimensional subspace F' of F are the roots of some nonzero linearized polynomial $\eta(x) = \sum_{i=0}^{b} \eta_i x^{p^i}$ over F (see [10, Ch. 4] or [11, Ch. 4]). The polynomial $\eta(x)$ defines a linear mapping $\eta : F \to F$ over $\mathrm{GF}(p)$. The range R_η of the mapping $x \mapsto \eta(x)$ over F is a subspace of F of dimension $h{-}b$ in which every two distinct elements have disjoint sets of p^b preimages under η. The set

$$\Big\{ \eta(x) - \beta \;:\; \beta \in R_\eta \Big\}$$

is thus simple over F.

4.3. The high-rate range: Proposition 1.4.

Proof of Proposition 1.4. We consider here codes at rates $(k{-}1)/n \ge 1{-}(2/(\ell{+}1)) = \rho_\ell$. Starting with the case $k = n$, we have $\ell < 2n \le 1{+}(q{-}1)n$ for all RS-admissible quadruples $(\ell, n, d, q) \ne (3, 2, 1, 2)$; so, in this case, $\Delta_\ell(n, 1; q) = 0$.

We assume from now on in the proof that $d = n{-}k{+}1$ is an even number (in such cases (6) holds); the case of odd d follows from Proposition 4.1(a). We show that there is an $[n, k, d]$ RS code $\mathcal{C}$ over $F = \mathrm{GF}(q)$ that contains a failing list of size $\ell{+}1$.

Let S_1' be the set $\{0, 1, \ldots, \ell\}$ and let $S_1, S_2, \ldots, S_{\ell+1}$ be the subsets of S_1' of size ℓ. Using any of the constructions of simple sets in section 4.2, we let $\{A_{S_i}(x)\}_{i=1}^{\ell+1}$ be a simple set over F, where $\deg A_{S_i}(x) = d/2$ for every i. We denote by U_i the set of $d/2$ roots of $A_{S_i}(x)$ in F and by U the union $\bigcup_{i=1}^{\ell+1} U_i$. Also, define U_1' to be a subset of $F \setminus U$ of size $n - (\ell{+}1)d/2$ and let

$$A_{S_1'}(x) = \prod_{\alpha \in U_1'} (x - \alpha) \;.$$

Define $\mathcal{P}$ to be a partition vector $(I_1, I_2, \ldots, I_{\ell+1}, I_1')$ of $\langle n \rangle$ with $|I_i| = d/2$ and $|I_1'| = n - (\ell{+}1)d/2$, and let $\mathcal{C}$ be defined by the code locators $\alpha_1, \alpha_2 \ldots, \alpha_n$, where $U_i = \{\alpha_\mu\}_{\mu \in I_i}$ and $U_1' = \{\alpha_\mu\}_{\mu \in I_1'}$.

By construction, $\mathcal{P}$ satisfies both (13) and (14) with equality. Finally, let the polynomials $f_{s,t}(x)$ be defined by (22). For every $s < t$, the set $\{\frac{f_{0,s}(x)}{g_{0,s,t}(x)}, \frac{f_{0,t}(x)}{g_{0,s,t}(x)}, \frac{f_{s,t}(x)}{g_{0,s,t}(x)}\}$ contains three different polynomials from $\{A_{S_i}(x)\}_{i=1}^{\ell+1}$ and is therefore a simple set over U. Hence, the polynomials $f_{0,s}(x)$, $f_{0,t}(x)$, and $f_{s,t}(x)$ satisfy the difference condition. Corollaries 3.5 and 3.6 now imply that $\mathcal{C}$ contains a failing list of size $\ell{+}1$. □

We next show that there are infinitely many quadruples satisfying the conditions of this proposition. The quadruples (ℓ, n, d, q), where $2 \le d \le 5$ and $\ell \le \frac{2n}{d} - 1$, satisfy the conditions for every field size $q \ge n$. Another example consists of the quadruples (ℓ, n, d, q), where $q = n = 2^m$, $d = 2^p$ for $1 < p < m$, and $\ell \le 2^{m-p+1} - 1$. In this case, $d/n = 2^{p-m} \le 2/(\ell + 1)$ and $\lceil (d-1)/2 \rceil = 2^{p-1}$ divides $q = 2^m$.

4.4. The midrate range. Propositions 4.2 and 4.3 below identify cases where the GS bound is tight for code rate around 0.3 and list sizes 3 or 4.

4.4.1. List-3 decoders for RS codes at rates ≈ 0.3.

PROPOSITION 4.2. *Let* $(3, 10m{+}\nu{+}\kappa, 7m{+}\nu{+}1, q)$ *be an RS-admissible quadruple, where* $q = p^h$ *for a prime* p*; the integer pair* (ν, κ) *belongs to the set* $\{(\nu,\kappa) : -1 \le \nu \le 1,\ 1 \le \kappa \le 5{-}3\nu\}$*; and* m *is a positive integer such that either* (a) $m \,|\, q{-}1$ *and* $q \ge 11m$ *or* (b) $m \,|\, q$ *and* $p \notin \{3,5,7\}$*. Then,*

$$\Delta_3^{RS}(10m{+}\nu{+}\kappa, 7m{+}\nu{+}1; q) = \lceil \tau_3(10m{+}\nu{+}\kappa, 7m{+}\nu{+}1) \rceil - 1 = 4m + \nu \ .$$

Proof. It suffices to prove the proposition for $(\nu,\kappa) = (-1,1)$, in which case $\tau_3(10m, 7m) = 4m$. The results for the remaining values of (ν,κ) follow from Proposition 4.1: for $\kappa = 1$, the result follows from part (a) of Proposition 4.1, and then, for every fixed ν, it follows from part (b). We construct an $[n{=}10m, k{=}3m{+}1, d{=}7m]$ RS code $\mathcal{C}$ over F that contains a failing list of size $\ell{+}1 = 4$; note that here $r = 2$ and that $\gamma = \gamma' = m$ satisfy (15).

Part (a). We assume that the field size q is such that $q{-}1 = m \cdot b$, where $b \ge 11$, and we let α be an element of order b in the multiplicative group of $F = \mathrm{GF}(q)$. We define six polynomials $A_{\{s,t\}}(x)$, $0 \le s < t \le 3$, and four polynomials $A_{\{s,t,u\}}(x)$, $0 \le s < t < u \le 3$, as follows:

$$\begin{array}{lll}
A_{\{0,1\}}(x) \ \ = x^m - \alpha^2 \ , & A_{\{0,2\}}(x) \ \ = x^m - \alpha \ , & A_{\{0,3\}}(x) \ \ = x^m - \alpha^7 \ , \\
A_{\{1,2\}}(x) \ \ = x^m - \alpha^3 \ , & A_{\{1,3\}}(x) \ \ = x^m - \alpha^9 \ , & A_{\{2,3\}}(x) \ \ = x^m - \alpha^8 \ , \\
A_{\{0,1,2\}}(x) = x^m - \alpha^{11} \ , & A_{\{0,1,3\}}(x) = x^m - \alpha^5 \ , & A_{\{0,2,3\}}(x) = x^m - \alpha^6 \ , \\
A_{\{1,2,3\}}(x) = x^m - \alpha^4 \ . & &
\end{array}$$

Note that any two of these ten polynomials are relatively prime, and each has m simple roots in F.

For every $S_i = \{s,t\}$ (respectively, $S'_j = \{s,t,u\}$), let U_i (respectively, U'_j) denote the set of roots of $A_{S_i}(x)$ (respectively, $A_{S'_j}(x)$) in F. Define accordingly a partition vector $\mathcal{P} = (I_i)_{i=1}^6 \| (I'_j)_{i=1}^4$ such that $U_i = \{\alpha_\mu\}_{\mu \in I_i}$ and $U'_j = \{\alpha_\mu\}_{\mu \in I'_j}$. Denote by U the union of $\bigcup_{i=1}^6 U_i$ and $\bigcup_{j=1}^4 U'_j$, and define $\mathcal{C}$ to be the $[10m, 3m{+}1, 7m]$ RS code over F whose set of code locators is U.

Next, we define the polynomials $f_{s,t}(x)$ by (22) and obtain

$$\begin{array}{ll}
f_{0,1}(x) = (x^m - \alpha^2)(x^m - \alpha^5)(x^m - \alpha^{11}), & f_{0,2}(x) = (x^m - \alpha)(x^m - \alpha^6)(x^m - \alpha^{11}), \\
f_{0,3}(x) = (x^m - \alpha^5)(x^m - \alpha^6)(x^m - \alpha^7), & f_{1,2}(x) = (x^m - \alpha^3)(x^m - \alpha^4)(x^m - \alpha^{11}), \\
f_{1,3}(x) = (x^m - \alpha^4)(x^m - \alpha^5)(x^m - \alpha^9), & f_{2,3}(x) = (x^m - \alpha^4)(x^m - \alpha^6)(x^m - \alpha^8).
\end{array}$$

Similarly, we define the polynomials $g_{0,s,t}(x)$ by (28), and the sets

$$\left\{ \tfrac{f_{0,s}(x)}{g_{0,s,t}(x)}, \tfrac{f_{0,t}(x)}{g_{0,s,t}(x)}, \tfrac{f_{s,t}(x)}{g_{0,s,t}(x)} \right\} \ , \quad 0 < s < t \le 3 \ , \tag{37}$$

are given by

$$\begin{aligned}
&\left\{ \tfrac{f_{0,1}(x)}{g_{0,1,2}(x)}, \tfrac{f_{0,2}(x)}{g_{0,1,2}(x)}, \tfrac{f_{1,2}(x)}{g_{0,1,2}(x)} \right\} \\
&\qquad = \left\{ (x^m{-}\alpha^2)(x^m{-}\alpha^5), (x^m{-}\alpha)(x^m{-}\alpha^6), (x^m{-}\alpha^3)(x^m{-}\alpha^4) \right\} , \\
&\left\{ \tfrac{f_{0,1}(x)}{g_{0,1,3}(x)}, \tfrac{f_{0,3}(x)}{g_{0,1,3}(x)}, \tfrac{f_{1,3}(x)}{g_{0,1,3}(x)} \right\} \\
&\qquad = \left\{ (x^m{-}\alpha^2)(x^m{-}\alpha^{11}), (x^m{-}\alpha^6)(x^m{-}\alpha^7), (x^m{-}\alpha^4)(x^m{-}\alpha^9) \right\} , \\
&\left\{ \tfrac{f_{0,2}(x)}{g_{0,2,3}(x)}, \tfrac{f_{0,3}(x)}{g_{0,2,3}(x)}, \tfrac{f_{2,3}(x)}{g_{0,2,3}(x)} \right\} \\
&\qquad = \left\{ (x^m{-}\alpha)(x^m{-}\alpha^{11}), (x^m{-}\alpha^5)(x^m{-}\alpha^7), (x^m{-}\alpha^4)(x^m{-}\alpha^8) \right\} .
\end{aligned}$$

Each of these three sets of polynomials is simple over U, since the three polynomials in each set differ only in their coefficient of x^m. Applying Corollary 3.6 to the partition vector $(I_i)_{i=1}^{6} \| (I'_j)_{i=1}^{4}$, it follows that $\mathcal{C}$ contains a failing list of size 4.

Part (b). We assume that the field size q is p^h for a prime $p \geq 11$ and that $m = p^b$ for $b < h$; the case $p = 2$ is omitted, as it is covered by Proposition 4.3 (to be proved immediately below). Let $\eta(x)$ be a linearized polynomial of degree m over $F = \mathrm{GF}(q)$ that has m simple roots in $\mathrm{GF}(q)$. Let β be a nonzero element in the range of the mapping $\eta : F \to F$; by linearity, the (distinct) elements $0, \beta, 2\beta, \ldots, 10\beta$ are also in that range. As in part (a), we define six polynomials $A_{\{s,t\}}(x)$, $0 \leq s < t \leq 3$, and four polynomials $A_{\{s,t,u\}}(x)$, $0 \leq s < t < u \leq 3$, each having m simple roots in F and every two being relatively prime:

$$\begin{array}{lll}
A_{\{0,1\}}(x) \;\; = \eta(x) - 2\beta \, , & A_{\{0,2\}}(x) \;\; = \eta(x) - \beta \, , & A_{\{0,3\}}(x) \;\; = \eta(x) - 7\beta \, , \\
A_{\{1,2\}}(x) \;\; = \eta(x) - 3\beta \, , & A_{\{1,3\}}(x) \;\; = \eta(x) - 9\beta \, , & A_{\{2,3\}}(x) \;\; = \eta(x) - 8\beta \, , \\
A_{\{0,1,2\}}(x) = \eta(x) - 11\beta \, , & A_{\{0,1,3\}}(x) = \eta(x) - 5\beta \, , & A_{\{0,2,3\}}(x) = \eta(x) - 6\beta \, , \\
A_{\{1,2,3\}}(x) = \eta(x) - 4\beta \, . & &
\end{array}$$

The proof now continues as in part (a); in particular, the sets (37) that result in this case are simple, as the three polynomials in each set differ only in their constant term. □

The failing list in Figure 1 is obtained from the construction in the proof of part (b) by taking $F = \mathrm{GF}(11)$, $m = 1$, $\eta(x) = x$, and $\beta = 1$.

4.4.2. List-4 decoders for RS codes at rates $\approx$ 0.3.

PROPOSITION 4.3. *Let* $(4, 10m{+}\nu{+}\kappa, 7m{+}\nu{+}1, q)$ *be an RS-admissible quadruple, where* (ν, κ) *is an integer pair in the set* $\{(2,1)\} \cup \{(\nu,\kappa) : -1 \leq \nu \leq 1,\ 1 \leq \kappa \leq 9{-}6\nu\}$ *and* q *and* m *are* powers of 2. *Then*

$$\Delta_4^{RS}(10m{+}\nu{+}\kappa, 7m{+}\nu{+}1; q) = \lceil \tau_4(10m{+}\nu{+}\kappa, 7m{+}\nu{+}1) \rceil - 1 = 4m + \nu \, .$$

Proof. We prove the proposition for $(\nu, \kappa) = (-1, 1)$, in which case $\tau_4(10m, 7m) = 4m$; the results for the other values for (ν, κ) extend directly from Proposition 4.1. We construct an $[n{=}10m, k{=}3m{+}1, d{=}7m]$ RS code $\mathcal{C}$ over $F = \mathrm{GF}(2^h)$ that contains a failing list of size $\ell{+}1 = 5$; here $r = 3$, and the values $\gamma = m$ and $\gamma' = 0$ satisfy (15).

Let $\eta(x)$ be a linearized polynomial of degree $m = 2^b \leq 2^{h-4}$ over F that has m simple roots in F. The range, R_η, of the mapping $x \mapsto \eta(x)$ over F is a linear space of dimension $h{-}b \geq 4$ over $\mathrm{GF}(2)$; therefore, one can find four elements $\beta_0, \beta_1, \beta_2, \beta_3 \in R_\eta$ that are linearly independent over $\mathrm{GF}(2)$. We represent each of the 16 elements $\sum_{i=0}^{3} \epsilon_i \beta_i$, $\epsilon_i \in \mathrm{GF}(2)$, as a 4-tuple $(\epsilon_0\, \epsilon_1\, \epsilon_2\, \epsilon_3)$.

Define the ten polynomials $A_{\{s,t,u\}}(x)$, $0 \leq s < t < u \leq \ell$, as follows:

$$\begin{array}{ll}
A_{\{0,1,2\}}(x) = \eta(x) - (1\,0\,0\,0) \, , & A_{\{0,1,3\}}(x) = \eta(x) - (0\,1\,0\,0) \, , \\
A_{\{0,1,4\}}(x) = \eta(x) - (0\,0\,1\,0) \, , & A_{\{0,2,3\}}(x) = \eta(x) - (0\,0\,0\,1) \, , \\
A_{\{0,2,4\}}(x) = \eta(x) - (0\,1\,1\,1) \, , & A_{\{0,3,4\}}(x) = \eta(x) - (1\,0\,1\,1) \, , \\
A_{\{1,2,3\}}(x) = \eta(x) - (1\,0\,0\,1) \, , & A_{\{1,2,4\}}(x) = \eta(x) - (1\,1\,1\,1) \, , \\
A_{\{1,3,4\}}(x) = \eta(x) - (0\,0\,1\,1) \, , & A_{\{2,3,4\}}(x) = \eta(x) - (0\,1\,1\,0) \, .
\end{array}$$

For every subset S_i of $\{0,1,2,3,4\}$ of size 3, let U_i denote the set of the $\gamma = m = 2^b$ roots of $A_{S_i}(x)$ in F, and denote by U the union $\bigcup_{i=1}^{10} I_i$. Define the partition vector $\mathcal{P} = (I_i)_{i=1}^{10}$ so that $U_i = \{\alpha_\mu\}_{\mu \in I_i}$. The code $\mathcal{C}$ is now defined as a $[10m, 3m{+}1, 7m]$ RS code over F whose set of code locators is U.

Let the polynomials $f_{s,t}(x)$ and $g_{s,t,u}(x)$ be defined by (22) and (28), respectively. It can be verified that each of the six sets

$$\left\{ \tfrac{f_{0,s}(x)}{g_{0,s,t}(x)}, \tfrac{f_{0,t}(x)}{g_{0,s,t}(x)}, \tfrac{f_{s,t}(x)}{g_{0,s,t}(x)} \right\}, \quad 0 < s < t \leq 4 , \tag{38}$$

is simple over F. In particular, the polynomials—each of degree $2m$—in every set differ only in their constant terms. For example,

$$\tfrac{f_{0,1}(x)}{g_{0,1,2}(x)} = A_{\{0,1,3\}}(x) \cdot A_{\{0,1,4\}}(x) \;=\; x^{2m} \;+\; (0\,1\,1\,0) \cdot x^m \;+\; (0\,1\,0\,0) \cdot (0\,0\,1\,0),$$
$$\tfrac{f_{0,2}(x)}{g_{0,1,2}(x)} = A_{\{0,2,3\}}(x) \cdot A_{\{0,2,4\}}(x) \;=\; x^{2m} \;+\; (0\,1\,1\,0) \cdot x^m \;+\; (0\,0\,0\,1) \cdot (0\,1\,1\,1),$$
$$\tfrac{f_{1,2}(x)}{g_{0,1,2}(x)} = A_{\{1,2,3\}}(x) \cdot A_{\{1,2,4\}}(x) \;=\; x^{2m} \;+\; (0\,1\,1\,0) \cdot x^m \;+\; (1\,0\,0\,1) \cdot (1\,1\,1\,1)$$

(multiplications are in F). Applying Corollary 3.6 to the proper partition vector $(I_i)_{i=1}^{10}$, it follows that $\mathcal{C}$ contains a failing list of size 5. □

The failing list in Figure 2 is obtained from the construction in the last proof by taking $F = \mathrm{GF}(2^4)$, $m = 1$, $\eta(x) = x$, and $\beta_i = \alpha^i$, where α is a root of $x^4 + x + 1$.

We turn next to proving Proposition 1.6, namely, to showing that the GS bound is not tight for RS-admissible quadruples $(\ell, n, d, q) = (4, 10, 7, q)$, where q is odd. The next lemma characterizes the structure of a failing list of size 5 in a $(10, M, 7)$ code over any field.

LEMMA 4.4. *Let $q = p^h$, where p is a prime. Every failing list of size 5 in a $(10, M, 7)$ code over $\mathrm{GF}(q)$ is a $(4, 3)$-configuration with respect to a proper partition vector $\mathcal{P} = (I_i)_{i=1}^{10}$, where $|I_i| = 1$ for all i. Every failing list of size 5 thus corresponds to a* BIBD(5, 3, 3).

Proof. The parameters $\ell = 4$, $r = 3$, $n = \binom{\ell+1}{r} = 10$, and $d = 7$ satisfy (6). Therefore, by Proposition 2.3, every failing list of size 5 forms a $(4, 3)$-configuration with respect to a partition vector $\mathcal{P} = (I_i)_{i=1}^{10} \| (I'_j)_{j=1}^5$ for which (13)–(14) hold with equality. Furthermore, since $\rho_r(\ell) = \rho_3(4) = 3/10 = 1 - d/n$, the partition vector $\mathcal{P}$ is proper; exactly $r = 3$ codewords agree on every position. We next show that each set I_i has size 1.

Assume the contrary; since $\sum_{i=1}^{10} |I_i| = 10$, at least one of the partition elements, say, I_1, is empty. Without loss of generality, let $S_1 = \{0, 1, 2\}$ and let the sets I_2 through I_7 correspond, respectively, to $S_2 = \{0, 1, 3\}$, $S_3 = \{0, 1, 4\}$, $S_4 = \{0, 2, 3\}$, $S_5 = \{0, 2, 4\}$, $S_6 = \{1, 2, 3\}$, and $S_7 = \{1, 2, 4\}$. We have

$$\sum_{i=2}^{7} |I_i| \;=\; \sum_{0 \leq s < t \leq 2} \Biggl(\sum_{i\,:\,\{s,t\} \subseteq S_i} |I_i| \Biggr) \;=\; 9 ,$$

where the second equality follows from the equality in (13). Hence, either $|I_2| + |I_4| + |I_6| \geq 5$ or $|I_3| + |I_5| + |I_7| \geq 5$. Assuming the former inequality (the arguments for the latter are similar) we obtain, again from (13),

$$\sum_{s \in \{0,1,2\}} \Biggl(\sum_{i\,:\,\{s,3\} \subseteq S_i} |I_i| \Biggr) \;\geq\; 2(|I_2| + |I_4| + |I_6|) \;\geq\; 10 .$$

Therefore, there must be $s \in \{0, 1, 2\}$ such that

$$\sum_{i\,:\,\{s,3\} \subseteq S_i} |I_i| \geq 4 ,$$

thereby contradicting (13). □

Proof of Proposition 1.6. Assume to the contrary that there is a $[10, 4, 7]$ RS code $\mathcal{C}$ over GF(q), q odd, that contains a failing list $\mathcal{L}$ of size 5. By Lemma 4.4, this failing list is a $(4, 3)$-configuration with respect to a proper partition vector $\mathcal{P} = (I_i)_{i=1}^{10}$, where $|I_i| = 1$ for all i. Let $\alpha_1, \alpha_2, \ldots, \alpha_{10}$ be the code locators of $\mathcal{C}$. The polynomials $A_{S_i}(x)$, which are defined by (21), can be written, without loss of generality, as

$$\begin{array}{lll}
A_{\{0,1,2\}}(x) = x - \alpha_1 \ , & A_{\{0,1,3\}}(x) = x - \alpha_2 \ , & A_{\{0,1,4\}}(x) = x - \alpha_3 \ , \\
A_{\{0,2,3\}}(x) = x - \alpha_4 \ , & A_{\{0,2,4\}}(x) = x - \alpha_5 \ , & A_{\{0,3,4\}}(x) = x - \alpha_6 \ , \\
A_{\{1,2,3\}}(x) = x - \alpha_7 \ , & A_{\{1,2,4\}}(x) = x - \alpha_8 \ , & A_{\{1,3,4\}}(x) = x - \alpha_9 \ , \\
A_{\{2,3,4\}}(x) = x - \alpha_{10} \ . & &
\end{array}$$

The polynomials $f_{s,t}(x)$, $0 \le s < t \le 4$, are defined accordingly by (22).

By Lemma 3.1(b), the ten polynomials $f_{s,t}(x)$ must satisfy (23). In particular, for every $0 \le s < t \le 4$, the three polynomials $f_{0,s}(x)/g_{0,s,t}$, $f_{0,t}(x)/g_{0,s,t}(x)$, and $f_{s,t}(x)/g_{0,s,t}(x)$, which take the form $(x - \alpha_{i_1})(x - \alpha_{i_2})$, must satisfy the difference condition. By Lemma 3.2, this happens if and only if the code locators satisfy the following six equations:

$$\begin{array}{ll}
(39) & (\alpha_1\alpha_3 - \alpha_7\alpha_9)(\alpha_1 + \alpha_3 - \alpha_4 - \alpha_6) = (\alpha_1\alpha_3 - \alpha_4\alpha_6)(\alpha_1 + \alpha_3 - \alpha_7 - \alpha_9), \\
(40) & (\alpha_1\alpha_2 - \alpha_8\alpha_9)(\alpha_1 + \alpha_2 - \alpha_5 - \alpha_6) = (\alpha_1\alpha_2 - \alpha_5\alpha_6)(\alpha_1 + \alpha_2 - \alpha_8 - \alpha_9), \\
(41) & (\alpha_2\alpha_3 - \alpha_7\alpha_8)(\alpha_2 + \alpha_3 - \alpha_4 - \alpha_5) = (\alpha_2\alpha_3 - \alpha_4\alpha_5)(\alpha_2 + \alpha_3 - \alpha_7 - \alpha_8), \\
(42) & (\alpha_2\alpha_6 - \alpha_7\alpha_{10})(\alpha_2 + \alpha_6 - \alpha_1 - \alpha_5) = (\alpha_2\alpha_6 - \alpha_1\alpha_5)(\alpha_2 + \alpha_6 - \alpha_7 - \alpha_{10}), \\
(43) & (\alpha_3\alpha_6 - \alpha_8\alpha_{10})(\alpha_3 + \alpha_6 - \alpha_1 - \alpha_4) = (\alpha_3\alpha_6 - \alpha_1\alpha_4)(\alpha_3 + \alpha_6 - \alpha_8 - \alpha_{10}), \\
 & (\alpha_2\alpha_4 - \alpha_9\alpha_{10})(\alpha_2 + \alpha_4 - \alpha_3 - \alpha_5) = (\alpha_2\alpha_4 - \alpha_3\alpha_5)(\alpha_2 + \alpha_4 - \alpha_9 - \alpha_{10}).
\end{array}$$

Defining

$$\begin{array}{ll}
 & \epsilon_7 = (\alpha_3 - \alpha_4)/(\alpha_7 - \alpha_4), \quad \epsilon_8 = (\alpha_2 - \alpha_5)/(\alpha_8 - \alpha_5), \\
(44) & \qquad \text{and} \quad \epsilon_9 = (\alpha_1 - \alpha_6)/(\alpha_9 - \alpha_6),
\end{array}$$

(39)–(41) can be rewritten as

$$(45) \qquad \begin{pmatrix} \alpha_2{-}\alpha_4 & \alpha_3{-}\alpha_5 & 0 \\ \alpha_1{-}\alpha_4 & 0 & \alpha_3{-}\alpha_6 \\ 0 & \alpha_1{-}\alpha_5 & \alpha_2{-}\alpha_6 \end{pmatrix} \begin{pmatrix} \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \end{pmatrix} = \begin{pmatrix} \alpha_2{-}\alpha_4{+}\alpha_3{-}\alpha_5 \\ \alpha_1{-}\alpha_4{+}\alpha_3{-}\alpha_6 \\ \alpha_1{-}\alpha_5{+}\alpha_2{-}\alpha_6 \end{pmatrix} .$$

Now, if the matrix in (45) were nonsingular, then the unique solution of (45) would be $\epsilon_7 = \epsilon_8 = \epsilon_9 = 1$, thereby requiring from (44) that certain code locators be identical, namely, $\alpha_7 = \alpha_3$, $\alpha_8 = \alpha_2$, and $\alpha_9 = \alpha_1$. Since this is impossible, the matrix in (45) must be singular, and this occurs if and only if

$$(46) \qquad -(\alpha_1 - \alpha_5)(\alpha_2 - \alpha_4)(\alpha_3 - \alpha_6) \ = \ (\alpha_1 - \alpha_4)(\alpha_2 - \alpha_6)(\alpha_3 - \alpha_5) \ .$$

Reiterating the analysis, with (39)–(41) now replaced by (41)–(43), we obtain

$$(47) \qquad -(\alpha_6 - \alpha_5)(\alpha_2 - \alpha_4)(\alpha_3 - \alpha_1) \ = \ (\alpha_6 - \alpha_4)(\alpha_2 - \alpha_1)(\alpha_3 - \alpha_5) \ .$$

Subtracting (46) from (47) and simplifying the result yield

$$2(\alpha_1 - \alpha_6)(\alpha_2 - \alpha_4)(\alpha_3 - \alpha_5) \ = \ 0 \ .$$

However, this is a contradiction whenever q is odd. We thus conclude that $\mathcal{C}$ cannot contain the failing list $\mathcal{L}$. $\square$

4.4.3. List-10 decoders for $[11, 3, 9]$ RS codes.

Proof of Proposition 1.7. Assume to the contrary that there is an $[11, 3, 9]$ RS code over $\mathrm{GF}(2^h)$ that contains a failing list $\mathcal{L}$ of size 11. By Proposition 2.3 and by property B1 in Proposition 2.8, the failing list corresponds to a symmetric BIBD$(11, 5, 2)$ (which has 11 blocks); namely, it forms a $(10, 5)$-configuration with respect to a proper partition vector $\mathcal{P} = (I_i)_i$ such that eleven partition elements I_i have size 1, whereas all the other partition elements in $\mathcal{P}$ are empty.

As this BIBD is essentially unique (see [2, p. 73]), we can assume, without loss of generality, that the nonempty partition elements in $\mathcal{P}$ are $I_i = \{i\}$, $1 \le i \le 11$, where $S_1, S_2, \ldots, S_{11}$ are given by

$$\begin{array}{lll} S_1 \ = \{1,3,4,5,9\}\ , & S_2 \ = \{2,4,5,6,10\}\ , & S_3 = \{0,3,5,6,7\}\ , \\ S_4 \ = \{1,4,6,7,8\}\ , & S_5 \ = \{2,5,7,8,9\}\ , & S_6 = \{3,6,8,9,10\}\ , \\ S_7 \ = \{0,4,7,9,10\}\ , & S_8 \ = \{0,1,5,8,10\}\ , & S_9 = \{0,1,2,6,9\}\ , \\ S_{10} = \{1,2,3,7,10\}\ , & S_{11} = \{0,2,3,4,8\}\ . & \end{array}$$

Define $A_{S_i}(x)$ and $f_{s,t}(x)$ accordingly by (21) and (22). In particular, we obtain

$$\begin{array}{ll} f_{0,2}(x) \ = (x-\alpha_9)(x-\alpha_{11})\ , & f_{0,7}(x) \ = (x-\alpha_3)(x-\alpha_7)\ , \\ f_{0,5}(x) \ = (x-\alpha_3)(x-\alpha_8)\ , & f_{0,10}(x) = (x-\alpha_7)(x-\alpha_8)\ , \\ f_{2,7}(x) \ = (x-\alpha_5)(x-\alpha_{10})\ , & f_{2,5}(x) \ = (x-\alpha_2)(x-\alpha_5)\ , \\ f_{2,10}(x) = (x-\alpha_2)(x-\alpha_{10})\ . & \end{array}$$

By Lemma 3.1(b), each of the following sets of three polynomials must satisfy the difference condition: $\{f_{0,2}(x), f_{0,7}(x), f_{2,7}(x)\}$, $\{f_{0,2}(x), f_{0,5}(x), f_{2,5}(x)\}$, and $\{f_{0,2}(x), f_{0,10}(x), f_{2,10}(x)\}$. By Lemma 3.2 we then obtain the following equations on the code locators:

$$(\alpha_9+\alpha_{11}-\alpha_3-\alpha_7)(\alpha_9\alpha_{11}-\alpha_5\alpha_{10}) = (\alpha_9+\alpha_{11}-\alpha_5-\alpha_{10})(\alpha_9\alpha_{11}-\alpha_3\alpha_7), \tag{48}$$

$$(\alpha_9+\alpha_{11}-\alpha_3-\alpha_8)(\alpha_9\alpha_{11}-\alpha_2\alpha_5) = (\alpha_9+\alpha_{11}-\alpha_2-\alpha_5)(\alpha_9\alpha_{11}-\alpha_3\alpha_8), \tag{49}$$

$$(\alpha_9+\alpha_{11}-\alpha_7-\alpha_8)(\alpha_9\alpha_{11}-\alpha_2\alpha_{10}) = (\alpha_9+\alpha_{11}-\alpha_2-\alpha_{10})(\alpha_9\alpha_{11}-\alpha_7\alpha_8). \tag{50}$$

Defining

$$\epsilon_3 = \frac{(\alpha_{11}-\alpha_3)(\alpha_9-\alpha_3)}{\alpha_5-\alpha_3}, \quad \epsilon_7 = \frac{(\alpha_{11}-\alpha_7)(\alpha_9-\alpha_7)}{\alpha_{10}-\alpha_7}, \quad \text{and} \quad \epsilon_8 = \frac{(\alpha_{11}-\alpha_8)(\alpha_9-\alpha_8)}{\alpha_2-\alpha_8},$$

we can rewrite (48)–(50) as

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \epsilon_3 \\ \epsilon_7 \\ \epsilon_8 \end{pmatrix} = \begin{pmatrix} \alpha_{11}-\alpha_3+\alpha_9-\alpha_7 \\ \alpha_{11}-\alpha_3+\alpha_9-\alpha_8 \\ \alpha_{11}-\alpha_8+\alpha_9-\alpha_7 \end{pmatrix}.$$

Summing up these equations and recalling that the field size is even, the left-hand side is identically zero, while the right-hand side equals the nonzero value $\alpha_9 + \alpha_{11}$, which is a contradiction. □

4.5. The low-rate range: Proposition 1.5.

Proof of Proposition 1.5. The proof is based on Lemma 3.7. One can verify that a sufficient condition for (33) to hold is that $\{\psi_{s,t}\}_{s,t}$ take either the form $\psi_{s,t} = \psi_s\psi_t$ or the form $\psi_{s,t} = \psi_s + \psi_t$ for some $\ell+1$ values $\psi_0, \psi_1, \ldots, \psi_\ell$. The values ψ_s must form a weak Sidon set (in the respective group) so as to have distinct values of $\psi_{s,t}$.

We now consider the two types of polynomials presented in section 4.2, taking e to be $k{-}1$.

Using polynomials of type 1 as the $\binom{\ell+1}{2}$ polynomials $\{f^*_{s,t}(x)\}_{0 \le s<t \le \ell}$ in Lemma 3.7, we require that $(k{-}1)|(q{-}1)$, and we select the respective constant terms $\psi_{0,1}, \psi_{0,2}, \ldots, \psi_{\ell-1,\ell}$ so that they satisfy $\psi_{s,t} = \psi_s \psi_t$. The set $\{\psi_0, \psi_1, \ldots, \psi_\ell\}$ should be a weak Sidon set of size $\ell{+}1$ in the multiplicative group of $\mathrm{GF}(q)$. If α is a primitive element in $\mathrm{GF}(q)$ and $\psi_s = \alpha^{\xi_s}$, then an equivalent requirement is that $\{\xi_0, \xi_1, \ldots, \xi_\ell\}$ be a weak Sidon set contained in the additive group of $\mathbb{Z}_{(q-1)/(k-1)}$.

When using polynomials of type 2 over $\mathrm{GF}(p^h)$ as $\{f^*_{s,t}(x)\}_{0 \le s<t \le \ell}$, we require that $k{-}1 = p^b$, where $b < h$, and we select the constant terms so that they satisfy $\psi_{s,t} = \psi_s + \psi_t$. The set $\{\psi_0, \psi_1, \ldots, \psi_\ell\}$ should be a weak Sidon set of size $\ell{+}1$ in the range, R_η, of a linearized polynomial $\eta(x)$ of degree p^b over F with p^b simple roots in $\mathrm{GF}(p^h)$. This range is an $(h{-}b)$-dimensional linear space over $\mathrm{GF}(p)$ and is therefore isomorphic to $\mathbb{Z}_p^{h-b}$. □

It is known that the additive group of $\mathbb{Z}_{(q-1)/(k-1)}$ contains a weak Sidon set of size $\ell{+}1$ whenever

$$\ell^2 \cdot (1 + o(1)) < (q{-}1)/(k{-}1) \,, \tag{51}$$

where $o(1)$ stands for an expression that goes to zero as $\ell \to \infty$ [6, Thm. 1]. In particular, for quadruples (ℓ, n, d, q), where $q = p^{2m}$, $n = q-1$, and $k = n-d+1 = p^m$, we get that $(k{-}1)|(q{-}1)$. The size of $\mathbb{Z}_{(q-1)/(k-1)}$ is $p^m + 1$. By [6, Thm. 1], $\mathbb{Z}_{(q-1)/(k-1)}$ contains some weak Sidon set of size $\ell{+}1$, where ℓ satisfies (51). It follows that $\ell^2 < n/(k-1)$, and thus $d/n \ge 1 - 2/(\ell(\ell+1))$, as required. We conclude that there are infinitely many quadruples that satisfy the requirements of Proposition 1.5 (part (a)).

For the group $\mathbb{Z}_p^{h-b}$ in part (b) of Proposition 1.5, the known bounds imply a weak Sidon set of size ℓ whenever

$$\ell^{2+o(1)} < q/(n{-}d) = p^{h-b} \tag{52}$$

(see [1, sect. 5]). If $n = q$, such a list size ℓ also satisfies the requirement $d/n \ge 1 - 2/(\ell(\ell+1))$. We can therefore find infinitely many quadruples $(\ell, n = p^h, d = p^h{-}p^b, q = p^h)$ satisfying the conditions of the proposition (part (b)).

Appendix.

Proof of Theorem 1.1. As shown in [7], the Guruswami–Sudan algorithm is a list-ℓ decoder with a decoding radius τ if there is a positive integer m such that the following two conditions hold:

$$r(n - \tau) \ge m + \ell(k{-}1) \text{ and} \tag{53}$$

$$\binom{r+1}{2} n < (\ell{+}1)m + \binom{\ell+1}{2}(k{-}1) \,. \tag{54}$$

(In terms of [7], $m + \ell(k{-}1)$ is the weighted degree of the bivariate polynomial $Q(x, y)$ which is computed by the algorithm.)

It can be easily verified that every integer τ that satisfies (53)–(54) for some positive integer m must be smaller than $\tau_\ell(n, d)$; therefore, $\tau \le \lceil \tau_\ell(n, d) \rceil - 1$. Next we show the converse result: we prove that $\tau = \lceil \tau_\ell(n, d) \rceil - 1$ satisfies (53)–(54) for some positive integer m. Define

$$m' = \tfrac{1}{\ell+1} \left(\binom{r+1}{2} n - \binom{\ell+1}{2}(k{-}1) \right) \,.$$

Since $(k-1)/n < \rho_{r+1}$, the value of m' is positive. We can now incorporate m' into the expression for $\tau_\ell(n,d)$ in (5) to obtain

$$(55)\quad \tau_\ell(n,d) = \tfrac{1}{(\ell+1)r}\left(\tbinom{\ell+1}{2}(n-(k-1)) - \tbinom{\ell+1-r}{2}n\right) = \tfrac{1}{r}\left(rn - m' - \ell(k-1)\right)\ .$$

If $\tau_\ell(n,d)$ is an integer, then m' must also be an integer; in this case, $m = m'+1$ and $\tau = \tau_\ell(n,d) - 1$ satisfy (53)–(54).

On the other hand, if $\tau_\ell(n,d)$ is not an integer, then

$$r\tau = r(\lceil \tau_\ell(n,d)\rceil - 1) < rn - m' - \ell(k-1)\ ,$$

and, therefore,

$$r\tau \le rn - \lceil m' \rceil - \ell(k-1)\ .$$

Hence, $m = \lceil m' \rceil$ and $\tau = \lceil \tau_\ell(n,d)\rceil - 1$ satisfy (53). Furthermore, this value of m is positive and satisfies (54) as well.

Fix the triple (ℓ, n, d), and consider the function

$$t_{\ell,n,d}(r) = \tfrac{1}{(\ell+1)r}\left(\tbinom{\ell+1}{2}d - \tbinom{\ell+1-r}{2}n\right)\ .$$

It can be easily verified that $t_{\ell,n,d}(r+1) \ge t_{\ell,n,d}(r)$ only for $1 - d/n \ge (r(r+1))/(\ell(\ell+1)) = \rho_{r+1}$. Hence, the value of r which maximizes the decoding radius of the Guruswami–Sudan algorithm is the one for which $1 - d/n = (k-1)/n \in [\rho_r, \rho_{r+1})$, as claimed in Theorem 1.1. $\square$

REFERENCES

[1] L. Babai and V. T. Sós, *Sidon sets in groups and induced subgraphs of Cayley graphs*, European J. Combin., 6 (1985), pp. 101–114.

[2] Th. Beth, D. Jungnickel, and H. Lenz, *Design Theory*, Cambridge University Press, Cambridge, UK, 1986.

[3] I. F. Blake and R. C. Mullin, *The Mathematical Theory of Coding*, Academic Press, New York, 1975.

[4] A. E. Brouwer, J. B. Shearer, N. J. A. Sloane, and W. D. Smith, *A new table of constant weight codes*, IEEE Trans. Inform. Theory, 36 (1990), pp. 1334–1380.

[5] O. Goldreich, R. Rubinfeld, and M. Sudan, *Learning polynomials with queries: The highly noisy case*, SIAM J. Discrete Math., 13 (2000), pp. 535–570.

[6] R. L. Graham and N. J. A. Sloane, *On additive bases and harmonious graphs*, SIAM J. Algebraic Discrete Methods, 1 (1980), pp. 382–404.

[7] V. Guruswami and M. Sudan, *Improved decoding of Reed-Solomon and algebraic-geometric codes*, IEEE Trans. Inform. Theory, 45 (1999), pp. 1757–1767.

[8] M. Hall, *Combinatorial Theory*, Wiley-Interscience, New York, 1986.

[9] J. Justesen and T. Høholdt, *Bounds on list decoding of MDS codes*, IEEE Trans. Inform. Theory, 47 (2001), pp. 1604–1609.

[10] R. Lidl and H. Niederreiter, *Finite Fields*, Addison-Wesley, Reading, MA, 1983.

[11] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North–Holland, Amsterdam, 1977.

[12] R. M. Roth and G. Seroussi, *Location-correcting codes*, IEEE Trans. Inform. Theory, 42 (1996), pp. 554–565.

[13] M. Sudan, *Decoding of Reed-Solomon codes beyond the error-correction bound*, J. Complexity, 13 (1997), pp. 180–193.

[14] I. Tal and R. M. Roth, *On list decoding of alternate codes in the Hamming and Lee metrics*, in Proceedings of the IEEE International Symposium on Information Theory (ISIT 2003), Yokohama, Japan, 2003, p. 364.

SIAM J. Discrete Math.
Vol. 17, No. 2, pp. 196–209

THE DETERMINATION OF THE CHAIN GOOD WEIGHT HIERARCHIES WITH HIGH DIMENSION*

YUAN LUO†, WENDE CHEN‡, AND A. J. HAN VINCK§

Abstract. There are a large number of linear block codes satisfying the chain condition. Their weight hierarchies are called chain good and form an important group in classifying all possible weight hierarchies. In this paper, we present a series of new sufficient conditions to determine which kinds of sequences are chain good weight hierarchies. Our results are efficient for the determination of the chain good weight hierarchies with high dimension.

Key words. generalized Hamming weight, weight hierarchy, chain condition

AMS subject classifications. 94B05, 94B65

DOI. 10.1137/S0895480102416684

1. Introduction.

The generalized Hamming weight and weight hierarchy were first introduced by Wei in [7] and Helleseth, Kløve, and Mykkeltveit in [4]. The rth generalized Hamming weight of a q-ary $[n, k]$ linear block code C is defined as

$$d_r = \min\{|\chi(D_r)| : D_r \text{ is an } [n, r] \text{ linear subcode of } C\},$$

where $\chi(D_r)$ is the support set of D_r, i.e.,

$$\chi(D_r) = \bigcup_{c \in D_r} \{e : c_e \neq 0, \text{ where } c = (c_1, \dots, c_n)\}.$$

The weight hierarchy of C is denoted by $(d_1, \dots, d_k)$. The chain condition was first introduced by Wei and Yang in [8]. We say that the code C satisfies the chain condition if there exist k subcodes $D_r(1 \leq r \leq k) \subseteq C$ such that

$$\dim(D_r) = r, \quad |\chi(D_r)| = d_r, \quad \text{and} \quad D_1 \subset D_2 \subset \cdots \subset D_k = C.$$

An integer sequence $(a_1, \dots, a_k)$ is called a "chain good weight hierarchy over $GF(q)$" if it is a weight hierarchy of an $[n, k]$ $(n = a_k)$ linear block code over $GF(q)$ satisfying the chain condition. In this paper, q is a fixed prime power. A chain good weight hierarchy over $GF(q)$ is also called a "chain good weight hierarchy."

There are a large number of linear block codes satisfying the chain condition; see [1, 2, 3, 5, 6, 8]. Their chain good weight hierarchies form an important group in classifying all possible weight hierarchies and they receive much attention. In [1] and [6], some sufficient conditions were given for the determination of the chain good weight hierarchies with general dimension over $GF(q)$. However, these conditions are not efficient for the determination of the chain good weight hierarchies with high

*Received by the editors October 24, 2002; accepted for publication (in revised form) May 19, 2003; published electronically November 4, 2003. This work was supported by National Science Foundation of China grant 10271116.

http://www.siam.org/journals/sidma/17-2/41668.html

†Computer Science & Engineering Department, Shanghai Jiao Tong University, Hua Shan Road 1954, Shanghai, 200030 China (yluo@cs.sjtu.edu.cn, jsluo@nankai.edu.cn).

‡Institute of Systems Science, Academy of Mathematics & System Sciences, Chinese Academy of Sciences, Beijing 100080, China (xinmei@public3.bta.net.cn).

§Institute for Experimental Mathematics, Duisburg-Essen University, Ellernstr. 29, 45326 Essen, Germany (vinck@exp-math.uni-essen.de).

dimension. In many cases, the lower bounds of these conditions increase exponentially with the dimension k; see the remarks of Theorems 2.1 and 2.2 in section 2.

In this paper, we present a series of new sufficient conditions to determine the chain good weight hierarchies with general dimension over $GF(q)$. The lower bounds of our new conditions increase linearly with the dimension k; see Corollaries 2.5 and 2.6 of section 2. They are more efficient than previous methods for the determination of the chain good weight hierarchies with high dimension.

Some preliminaries and our main results are introduced in section 2. In section 3, some interesting properties are shown. The proofs of our main results are presented in sections 4 and 5. For $q = 3$ and $k = 6, 7, 8$, the improvements on [1] and [6] are listed in section 6. Section 7 is the conclusion.

2. Preliminaries and main results. A positive integer sequence $(a_1, \ldots, a_k)$ is called chain permissible over $GF(q)$ if $qi_{r-1} \geq i_r \geq 0$ for $1 \leq r \leq k-1$, where

$$i_r = a_{k-r} - a_{k-r-1} \quad \text{and} \quad a_0 = 0. \tag{2.1}$$

We know that the chain good weight hierarchies are chain permissible [3] and there also exist some chain permissible sequences which do not correspond to any weight hierarchies [2]. From (2.1), it is easy to see that the parameter sequence $(i_0, \ldots, i_{k-1})$ can be determined from the sequence $(a_1, \ldots, a_k)$ and vice versa. Let

$$\pi_r = (1-q)\sum_{j=0}^{r-1} i_j + i_r = \sum_{j=1}^{r}(i_j - qi_{j-1}) + i_0 \quad \text{for} \quad 0 \leq r \leq k-1. \tag{2.2}$$

Then

$$i_r = \sum_{j=0}^{r} \pi_j S_{r,j} \quad \text{for} \quad 0 \leq r \leq k-1, \tag{2.3}$$

where $S_{j,l} = (q-1)q^{j-l-1}$ for $j > l$, and $S_{j,j} = 1$. For a chain permissible sequence $(a_1, \ldots, a_k)$, it is easy to see from (2.2) that

$$\pi_0 \geq \cdots \geq \pi_{k-1}. \quad (\pi_r \text{ may be negative for } r \geq 1.) \tag{2.4}$$

Denote

$$\iota_r = \lfloor i_r/q^r \rfloor \quad \text{and} \quad p_r = i_r - \iota_r q^r \quad \text{for} \quad 0 \leq r \leq k-1 \tag{2.5}$$

and

$$\delta_r = \begin{cases} 0 & \text{if} \quad 0 \leq p_{r+1} \leq p_r q, \\ 1 & \text{if} \quad p_r q < p_{r+1} < q^{r+1}. \end{cases} \tag{2.6}$$

Then for any chain permissible sequence, it was shown in [1] that

$$\iota_r \geq \iota_{r+1} + \delta_r. \tag{2.7}$$

The following theorems, Theorems 2.1 [1] and 2.2 [6], are two methods for the determination of the chain good weight hierarchies.

THEOREM 2.1 (see [1]). *A chain permissible sequence* $(a_1, \ldots, a_k)$ *is a chain good weight hierarchy if*

$$\iota_{k-3} \geq (q-1) + \sum_{r=0}^{k-4}(\delta_r(q^{r+1}-1) + qp_r - p_{r+1}). \tag{2.8}$$

Remark. For fixed q, it is easy to see that the lower bound of condition (2.8) is greater than $q^{k-3}/2$ if $\delta_{k-4} = 1$ and $p_{k-3} - qp_{k-4}$ is small positive; it is also greater than $q^{k-4}/2$ if $\delta_{k-5} = 1$ and $p_{k-4} - qp_{k-5}$ is small positive, and so on. Therefore, in many cases, the lower bound of the condition (2.8) increases exponentially with the dimension k. By using (2.7) for a chain permissible sequence, we know that the exponential increase of ι_{k-3} with k implies the exponential increase of ι_r with k for $0 \leq r \leq k-4$.

Theorem 2.2 (see [6]). *A chain permissible sequence $(a_1, \ldots, a_k)$ is a chain good weight hierarchy if*

$$\iota_{k-2} \geq \sum_{r=0}^{k-3} (\delta_r(q^{r+1} - 1) + qp_r - p_{r+1}). \tag{2.9}$$

Remark. By the same arguments as in the remark for Theorem 2.1, we know that the lower bound of condition (2.9) also increases exponentially with the dimension k in many cases. The exponential increase of ι_{k-2} with k implies the exponential increase of ι_r with k for $0 \leq r \leq k-3$.

Therefore, Theorems 2.1 and 2.2 are not so efficient for large k. In this paper, we present a series of new sufficient conditions, the lower bounds of which increase linearly with the dimension k; see Corollaries 2.5 and 2.6. These new conditions are more efficient for the determination of the chain good weight hierarchies with high dimension. The following theorem provides an original idea about how to give a sufficient condition by using the parameters $\pi_0, \ldots, \pi_\Gamma$, where $0 \leq \Gamma \leq k-2$.

Theorem 2.3. *For a chain permissible sequence $(a_1, \ldots, a_k)$ and an integer Γ such that $0 \leq \Gamma \leq k-2$, if there exist some integers $\theta_0 \geq \theta_1 \geq \cdots \geq \theta_{k-2} \geq 0$ satisfying*

$$i_{k-2} = \sum_{l=0}^{k-2} \theta_l S_{k-2,l}, \quad \textit{where} \quad 0 \leq \theta_l \leq \pi_l \quad \textit{for} \quad 0 \leq l \leq \Gamma, \tag{2.10}$$

and

$$i_{j-1} \geq i_j/q + S_{j-1,0} \quad \textit{for} \quad \Gamma + 2 \leq j \leq k-3, \tag{2.11}$$

then the chain permissible sequence is a chain good weight hierarchy.

In Theorem 2.3, condition (2.11) does not exist for $\Gamma = k-2$, $k-3$, and $k-4$. For $\Gamma = k-2$, the corresponding result of Theorem 2.3 was obtained in [6]; i.e., a chain permissible sequence $(a_1, \ldots, a_k)$ is a chain good weight hierarchy if $\pi_{k-2} \geq 0$. For $\Gamma = k-4$, the corresponding result of Theorem 2.3 includes the cases where $\Gamma = k-2$ and $k-3$.

Note that the integers $\theta_0, \ldots, \theta_{k-2}$ satisfying (2.10) *do not exist if $\pi_\Gamma < 0$.* In fact, if π_Γ and ι_{k-2} are large positive, we can find some suitable $\theta_0, \ldots, \theta_{k-2}$ and get the following theorem.

Theorem 2.4. *Let $(a_1, \ldots, a_k)$ be a chain permissible sequence and let Γ be an integer such that $0 \leq \Gamma \leq k-4$. Then $(a_1, \ldots, a_k)$ is a chain good weight hierarchy if*

$$\pi_\Gamma \geq (k-2)q, \quad \iota_{k-2} \geq (k-2)(q-1), \tag{2.12}$$

and

$$i_{j-1} \geq i_j/q + S_{j-1,0} \quad \textit{for} \quad \Gamma + 2 \leq j \leq k-3. \tag{2.13}$$

Furthermore, $(a_1, \ldots, a_k)$ is a chain good weight hierarchy if

$$\iota_\Gamma \geq (k-2)q + \sum_{r=0}^{\Gamma-1} (\delta_r (q^{r+1}-1) + qp_r - p_{r+1}), \tag{2.14}$$

$$\iota_{k-2} \geq (k-2)(q-1), \tag{2.15}$$

and

$$i_{j-1} \geq i_j / q + S_{j-1,0} \quad \textit{for} \quad \Gamma + 2 \leq j \leq k-3. \tag{2.16}$$

Note that, when $\Gamma = k-4$, (2.13) and (2.16) do not exist.

Theorem 2.4 presents a series of new sufficient conditions by using a different Γ. A large number of new chain good weight hierarchies are found by using this theorem (see section 6). In particular, from the second part of Theorem 2.4, we get two important results, Corollaries 2.5 and 2.6.

COROLLARY 2.5. *Let Γ be a fixed nonnegative integer. A chain permissible sequence $(a_1, \ldots, a_k)$, where $k \geq \Gamma + 6$, is a chain good weight hierarchy if*

$$\iota_{k-2} \geq (k-2)(q-1) + \sum_{r=0}^{\Gamma-1} q^{r+1} \quad \textit{and} \tag{2.17}$$

$$\iota_{j-1} \geq \iota_j + 2 \quad \textit{for} \quad \Gamma + 2 \leq j \leq k-3. \tag{2.18}$$

Remark. In Corollary 2.5, the lower bound of condition (2.17) increases linearly with the dimension k. The linear increase of ι_{k-2} with k only implies the linear increase of ι_r with k for $0 \leq r \leq k-3$. Therefore, in the determination of the chain good weight hierarchies with high dimension, Corollary 2.5 is more efficient than Theorems 2.1 and 2.2. By the same arguments, we have Corollary 2.6. In Corollary 2.6, the lower bound on the condition for ι_{k-2} is smaller, but a larger k is needed.

COROLLARY 2.6. *Let Γ be a fixed nonnegative integer. A chain permissible sequence $(a_1, \ldots, a_k)$, where $k \geq \sum_{r=0}^{\Gamma-1} q^{r+1} + 6$, is a chain good weight hierarchy if*

$$\iota_{k-2} \geq (k-2)(q-1) \quad \textit{and} \tag{2.19}$$

$$\iota_{j-1} \geq \iota_j + 2 \quad \textit{for} \quad \Gamma + 2 \leq j \leq k-3. \tag{2.20}$$

3. Some basic lemmas. In this section, we give some interesting properties, which are useful in establishing our main results. In section 3.1, two types of expressions are introduced. We show that a nonnegative integer having a type I expression can also be expressed in type II. Then, in section 3.2, a symbol $R(\cdot, \cdot)$ is used to describe the relation of two expressions. In the last subsection, we introduce two new parameters, π_j^* and T_j, of a chain permissible sequence.

3.1. Two types of expressions. For nonnegative integers $z_0, \ldots, z_J$, let $[z_0, \cdots, z_J]$ be the expression $\sum_{l=0}^{J} z_l S_{J,l}$, where $J \geq 1$. We say that $[z_0, \ldots, z_J] = [y_0, \ldots, y_J]$ if $z_l = y_l$ for $0 \leq l \leq J$. We say that $[z_0, \ldots, z_J]$ and $[y_0, \ldots, y_J]$ have the same value if $\sum_{l=0}^{J} z_l S_{J,l} = \sum_{l=0}^{J} y_l S_{J,l}$. Let

$$D[z_0, \ldots, z_J] = [z_0 - \Delta, z_1 + \Delta(q-1), \ldots, z_J + \Delta(q-1)], \tag{3.1}$$

where $\Delta = \lfloor \frac{z_0 - z_1}{q} \rfloor$. Then $D[z_0, \ldots, z_J]$ and $[z_0, \ldots, z_J]$ have the same value. The expression $[z_0, \ldots, z_J]$ is called type I if

$$z_l \geq z_{l+1} \geq 0 \quad \text{for all} \;\; 0 \leq l \leq J-1. \tag{3.2}$$

It is called type II if

$$z_{l+1} + q > z_l \geq z_{l+1} \geq 0 \quad \text{for all } 0 \leq l \leq J-1. \tag{3.3}$$

Furthermore, we have the following property.

Lemma 3.1. *Let $[z_0, \ldots, z_J]$ be an expression of type* I; *then an expression of type* II *having the same value can be given by $[z_0^{(J)}, \ldots, z_J^{(J)}]$, where*

$$[z_{J-l}^{(l)}, \ldots, z_J^{(l)}] = D[z_{J-l}, z_{J-l+1}^{(l-1)}, \ldots, z_J^{(l-1)}] \textit{ for } 1 \leq l \leq J, \ \ z_J^{(0)} = z_J. \tag{3.4}$$

Proof. For $l = 1$, it is easy to see that $[z_{J-1}^{(1)}, z_J^{(1)}] = D[z_{J-1}, z_J]$ is type II. For $l = t$, suppose $[z_{J-t}^{(t)}, \ldots, z_J^{(t)}]$ is type II. Then for $l = t+1$, the expression

$$[z_{J-t-1}^{(t+1)}, \ldots, z_J^{(t+1)}] = D[z_{J-t-1}, z_{J-t}^{(t)}, \ldots, z_J^{(t)}]$$

is also type II. Therefore, by induction, $[z_0^{(J)}, \ldots, z_J^{(J)}]$ is an expression of type II. Furthermore, $[z_0^{(J)}, \ldots, z_J^{(J)}]$ and $[z_0, z_1, \ldots, z_J]$ have the same value since the operator D does not change the value of an expression. □

3.2. A relation R of two expressions. Let SUM_j and SUM_{j+1} be two expressions such that

$$SUM_j : \sum_{l=0}^{j} \alpha_{j,l} S_{j,l} + \lambda_{j,l} \quad \text{and} \quad SUM_{j+1} : \sum_{l=0}^{j+1} \alpha_{j+1,l} S_{j+1,l} + \lambda_{j+1,l}, \tag{3.5}$$

where $\alpha_{j,l}, \alpha_{j+1,l}, \lambda_{j,l} (< S_{j,l})$, and $\lambda_{j+1,l} (< S_{j+1,l})$ are nonnegative integers. We say that

$$R(SUM_j, SUM_{j+1}) \text{ is true} \tag{3.6}$$

if the coefficients of SUM_j and SUM_{j+1} satisfy

$$\alpha_{j,l} \geq \alpha_{j,l+1} + \epsilon(\lambda_{j,l+1}), \tag{3.7}$$

$$\alpha_{j+1,l} \geq \alpha_{j+1,l+1} + \epsilon(\lambda_{j+1,l+1}), \tag{3.8}$$

$$\alpha_{j,l} \geq \alpha_{j+1,l} + \epsilon(\lambda_{j+1,l}), \tag{3.9}$$

where $\epsilon(x) = 0$ for $x = 0$ and $\epsilon(x) = 1$ otherwise. By using the symbol $R(\cdot, \cdot)$, Theorem 2 of [6] can be given as follows.

Lemma 3.2 (see [6]). *For a chain permissible sequence $(a_1, \ldots, a_k)$, if there exist nonnegative integers $\alpha_{j,l}$ and $\lambda_{j,l} (< S_{j,l})$ such that*

$$E_j : \quad i_j = \sum_{l=0}^{j} \alpha_{j,l} S_{j,l} + \lambda_{j,l} \quad \textit{for} \quad 0 \leq j \leq k-1 \textit{ and} \tag{3.10}$$

$$R(E_j, E_{j+1}) \textit{ is true for } 0 \leq j \leq k-2, \tag{3.11}$$

then it is a chain good weight hierarchy.

3.3. New parameters: π_j^* and T_j. For a chain permissible sequence $(a_1,\dots,a_k)$, the relation between the parameter sequences $(i_0,\dots,i_{k-1})$ and $(\pi_0,\dots,\pi_{k-1})$ is obtained in (2.2) and (2.3). Now, we introduce a new parameter sequence $(\pi_0^*,\dots,\pi_{k-1}^*)$, which is useful for studying the bound of i_j $(0\le j\le k-1)$. For $0\le\Gamma\le k-4$, let

$$\pi_l^*=\pi_l \quad\text{for}\quad 0\le l\le\Gamma \quad\text{and}\quad \pi_l^*=\pi_\Gamma \quad\text{for}\quad \Gamma+1\le l\le k-1. \tag{3.12}$$

Denote

$$T_j=\sum_{l=0}^{j}\pi_l^*S_{j,l} \quad\text{for}\quad 0\le j\le k-1. \tag{3.13}$$

Lemma 3.3. *For a chain permissible sequence $(a_1,\dots,a_k)$, we have*

$$i_j\le T_j \quad \textit{for}\quad 0\le j\le k-1. \tag{3.14}$$

If $i_{\Gamma+1}>i_{\Gamma+2}/q$, we have

$$i_j< T_j \quad \textit{for}\quad j\ge\Gamma+2. \tag{3.15}$$

Proof. For a chain permissible sequence $(a_1,\dots,a_k)$, it is shown in (2.4) that $\pi_0\ge\cdots\ge\pi_{k-1}$. Then $\pi_l\le\pi_l^*$ for $0\le l\le k-1$ and

$$i_j=\sum_{l=0}^{j}\pi_lS_{j,l}\le T_j \quad\text{for}\quad 0\le j\le k-1.$$

When $i_{\Gamma+1}>i_{\Gamma+2}/q$, if there exists an integer $j\ge\Gamma+2$ such that $i_j=T_j$, then

$$\begin{aligned}
&i_j=\sum_{l=0}^{j}\pi_lS_{j,l}=\sum_{l=0}^{j}\pi_l^*S_{j,l}\\
\Rightarrow&\pi_j=\pi_\Gamma\\
\Rightarrow&\sum_{t=\Gamma+1}^{j}(i_t-qi_{t-1})=0\\
\Rightarrow&i_{t-1}=i_t/q \quad\text{for}\quad \Gamma+1\le t\le j,
\end{aligned}$$

which is impossible. □

4. Proof of Theorem 2.3. In this section, the proof of Theorem 2.3 is given in two parts. The first part is presented for $\Gamma=k-4$ in Lemma 4.2, i.e., Theorem 4 of [6]. Now, we have a new description of the proof, which is useful in establishing the whole proof of Theorem 2.3. The second part is presented for $\Gamma\le k-5$. In addition, the following lemma, which is derived from Lemma 5 of [2], allows us to pay attention only to some special chain permissible sequences satisfying $i_{k-1}=qi_{k-2}$.

Lemma 4.1 (see [2]). *For fixed integers $i_0^*,\dots,i_{k-2}^*$, let $\mathcal{A}$ be the set of chain permissible sequences with dimension k such that $i_j=i_j^*$ $(0\le j\le k-2)$. Then all of the sequences in $\mathcal{A}$ are chain good weight hierarchies if the sequence in $\mathcal{A}$ satisfying $i_{k-1}=qi_{k-2}$ is a chain good weight hierarchy.*

LEMMA 4.2 (see [6]). *For a chain permissible sequence* $(a_1, \ldots, a_k)$*, if there exist some integers* $\theta_0 \geq \theta_1 \geq \cdots \geq \theta_{k-2} \geq 0$ *such that*

$$i_{k-2} = \sum_{l=0}^{k-2} \theta_l S_{k-2,l}, \quad \textit{where} \quad \theta_l \leq \pi_l \quad \textit{for} \quad 0 \leq l \leq k-4, \tag{4.1}$$

then it is a chain good weight hierarchy.

Proof. By using Lemmas 3.1 and 4.1, we can assume that $[\theta_{k-3}, \theta_{k-2}]$ is type II and $i_{k-1} = qi_{k-2}$. Since

$$E_j : \qquad i_j = \sum_{l=0}^{j} \pi_l S_{j,l} \quad \text{for} \quad 0 \leq j \leq k-4,$$

$$E_{k-2} : \qquad i_{k-2} = \sum_{l=0}^{k-2} \theta_l S_{k-2,l},$$

$$E_{k-1} : \qquad i_{k-1} = qi_{k-2} = \sum_{l=0}^{k-2} \theta_l S_{k-1,l} + \theta_{k-2},$$

it follows that this lemma can be obtained by using Lemma 3.2 if there exists a suitable expression E_{k-3} for i_{k-3} such that $R(E_{k-4}, E_{k-3})$ and $R(E_{k-3}, E_{k-2})$ are both true.

In the following paragraphs, after showing two bounds of i_{k-3}, a suitable expression E_{k-3} is given in (4.4). The first bound is an upper bound obtained from Lemma 3.3:

$$i_{k-3} \leq T_{k-3} = \sum_{l=0}^{k-3} \pi_l^* S_{k-3,l}. \tag{4.2}$$

The second bound is a lower bound. Denote $\Lambda = \sum_{l=0}^{k-3} \theta_l S_{k-3,l}$; we have

$$i_{k-3} \geq \lceil i_{k-2}/q \rceil = \Lambda \tag{4.3}$$

since $i_{k-3} \geq i_{k-2}/q$ and $[\theta_{k-3}, \theta_{k-2}]$ is type II. Then a suitable expression E_{k-3} for i_{k-3} is obtained in (4.4), where the coefficients are less than or equal to those of T_{k-3} and greater than or equal to those of Λ. Denote

$$e_l = \pi_l^* - \theta_l \quad \text{for} \quad 0 \leq l \leq k-3,$$

$$L = \max \left\{ \delta : i_{k-3} \geq \Lambda + \sum_{l=0}^{\delta} e_l S_{k-3,l} \right\} \quad (\text{let } L = -1 \text{ if } \delta \text{ does not exist}),$$

$$g = i_{k-3} - \Lambda - \sum_{l=0}^{L} e_l S_{k-3,l};$$

we have

$$\begin{aligned} i_{k-3} &= \Lambda + \sum_{l=0}^{L} e_l S_{k-3,l} + g \\ &= \sum_{l=0}^{L} \pi_l^* S_{k-3,l} + ((\theta_{L+1} + g_1) S_{k-3,L+1} + g_2) + \sum_{l=L+2}^{k-3} \theta_l S_{k-3,l}, \end{aligned} \tag{4.4}$$

where $g_1 = \lfloor g/S_{k-3,L+1} \rfloor < e_{k-3}$ and $g_2 = g - g_1 S_{k-3,L+1} < S_{k-3,L+1}$. For $L = k-4$, the last part of (4.4) does not exist. For $L = k-3$, the last two parts of (4.4) do not exist. □

Proof of Theorem 2.3 *when* $\Gamma \leq k-5$. By using Lemmas 3.1 and 4.1, we can assume that $[\theta_{\Gamma+1}, \ldots, \theta_{k-2}]$ is type II and $i_{k-1} = q i_{k-2}$.

From Lemma 3.2, we know that this theorem can be obtained if there exist the following expressions:

$$E_j : \; i_j = \sum_{l=0}^{j} \pi_l^* S_{j,l} \quad \text{for} \quad 0 \leq j \leq \Gamma, \tag{4.5}$$

$$E_j : \; i_j = \sum_{l=0}^{u_j-1} \pi_l^* S_{j,l} + \sum_{l=u_j}^{j} \alpha_{j,l} S_{j,l} + \lambda_{j,\eta_j} \text{ for } \Gamma+1 \leq j \leq k-3, \tag{4.6}$$

$$E_{k-2} : \; i_{k-2} = \sum_{l=0}^{k-2} \theta_l S_{k-2,l},$$

$$E_{k-1} : \; i_{k-1} = \sum_{l=0}^{k-2} \theta_l S_{k-1,l} + \theta_{k-2},$$

where $\alpha_{j,l}$, u_j, $\eta_j(\geq u_j)$, and $\lambda_{j,\eta_j}(< S_{j,\eta_j})$ are nonnegative integers to be determined under the true condition $R(E_j, E_{j+1})$. Note that expression (4.5) is fixed.

In the following paragraphs, the construction for (4.6) is given in three steps. In Step 1, an expression E_j is obtained from E_{j+1} by induction in (4.8). Then, in Step 2, we show that $R(E_j, E_{j+1})$ is true. However, in some cases, E_j should be changed. The changes are given in the last step.

Step 1. Now, we show how to get the expression (4.6) by induction. By the same arguments as in the proof of Lemma 4.2, we get an expression E_{k-3} from E_{k-2} such that

$$R(E_{k-3}, E_{k-2}) \text{ is true and } u_{k-3} = \eta_{k-3}.$$

For any integer $j : \Gamma+1 \leq j \leq k-4$, assume that E_{j+1} has been obtained from E_{j+2} satisfying

$$R(E_{j+1}, E_{j+2}) \text{ is true and } u_{j+1} = \eta_{j+1}.$$

Then, by the same arguments as in the proof of Lemma 4.2, we get an expression E_j in (4.8) from E_{j+1} if

$$[\alpha_{j+1,u_{j+1}^*}, \ldots, \alpha_{j+1,j+1}] \text{ is type II,} \tag{4.7}$$

where $u_{j+1}^* = \max\{u_{j+1}, \Gamma+1\}$. The corresponding arguments are

$$\Lambda = \sum_{l=0}^{u_{j+1}-1} \pi_l^* S_{j,l} + \sum_{l=u_{j+1}}^{j} \alpha_{j+1,l} S_{j,l} \leq \lceil i_{j+1}/q \rceil \leq i_j \quad \text{(by using (4.7))},$$

$$e_l = \pi_l^* - \alpha_{j+1,l} \quad \text{for} \quad u_{j+1} \leq l \leq j,$$

$$L = \max\left\{ \delta : i_j \geq \Lambda + \sum_{l=u_{j+1}}^{\delta} e_l S_{j,l} \right\} \quad \text{(if } \delta \text{ doesn't exist, let } L = u_{j+1} - 1\text{)},$$

$$g = i_j - \Lambda - \sum_{l=u_{j+1}}^{L} e_l S_{j,l}.$$

Denote $g_1 = \lfloor g/S_{j,L+1} \rfloor$ and $g_2 = g - g_1 S_{j,L+1}$; we have

$$
\begin{aligned}
i_j &= \Lambda + \sum_{l=u_{j+1}}^{L} e_l S_{j,l} + g \\
&= \sum_{l=0}^{L} \pi_l^* S_{j,l} + ((a_{j+1,L+1} + g_1) S_{j,L+1} + g_2) + \sum_{l=L+2}^{j} a_{j+1,l} S_{j,l} \\
&= \sum_{l=0}^{L} \pi_l^* S_{j,l} + \sum_{l=L+1}^{j} \alpha_{j,l} S_{j,l} + \lambda_{j,L+1}, \qquad (4.8)
\end{aligned}
$$

where $\alpha_{j,L+1} = \alpha_{j+1,L+1} + g_1$, $\alpha_{j,l} = \alpha_{j+1,l}$ for $L+2 \le l \le j$ and $\lambda_{j,L+1} = g_2 < S_{j,L+1}$. Note that in (4.8) the coefficients are greater than or equal to those of Λ and less than or equal to those of T_j. In addition,

$$u_j = \eta_j = L + 1 \ge u_{j+1} = \eta_{j+1}. \qquad (4.9)$$

Step 2. By analyzing two cases of (4.9), we know that $R(E_j, E_{j+1})$ is true.

- If $L + 1 > u_{j+1}$, then it is easy to verify that $R(E_j, E_{j+1})$ is true.
- Assume that $L+1 = u_{j+1}$. By using (2.11), we have $i_j - i_{j+1}/q \ge S_{j,0}$. Then $g = i_j - \Lambda \ge i_j - \lceil i_{j+1}/q \rceil \ge S_{j,0}$ and

  $$g_1 \ge \lfloor S_{j,0}/S_{j,L+1} \rfloor \ge 1, \qquad (4.10)$$

  which implies that $R(E_j, E_{j+1})$ is true.

Step 3. In Step 1, we construct E_j from E_{j+1} by induction when E_{j+1} satisfies (4.7). For E_{k-2}, condition (4.7) is obvious since $[\theta_{\Gamma+1}, \dots, \theta_{k-2}]$ is type II. Now we should make E_j have the same property, where $\Gamma + 2 \le j \le k - 3$.

Suppose E_{j+1} has property (4.7), and E_j is obtained in Step 1. In the following two cases, we present a method to make $[\alpha_{j,u_j^*}, \dots, \alpha_{j,j}]$ a type II expression. Note that u_j^* denotes $\max\{\mu_j, \Gamma + 1\}$.

- *Case* 1. $u_j^* < j$.
  - If $u_j < \Gamma+1$, then, from (4.8) and (4.9), we know that $u_j^* = \Gamma+1 = u_{j+1}^*$, and $[\alpha_{j,u_j^*}, \dots, \alpha_{j,j}] = [\alpha_{j+1,u_{j+1}^*}, \dots, \alpha_{j+1,j}]$ is type II.
  - If $u_j \ge \Gamma + 1$, then $u_j^* = u_j$. Let

    $$[\alpha'_{j,u_j}, \dots, \alpha'_{j,j}] = D[\alpha_{j,u_j}, \dots, \alpha_{j,j}].$$

    Then it is easy to verify that $[\alpha'_{j,u_j}, \dots, \alpha'_{j,j}]$ is type II since $[\alpha_{j,u_j+1}, \dots, \alpha_{j,j}] = [\alpha_{j+1,u_j+1}, \dots, \alpha_{j+1,j}]$ is type II. Now, we get a new expression for i_j:

    $$E'_j: \quad i_j = \sum_{l=0}^{u_j-1} \pi_l^* S_{j,l} + \sum_{l=u_j}^{j} \alpha'_{j,l} S_{j,l} + \lambda_{j,u_j}. \qquad (4.11)$$

    E_j can be replaced with E'_j since $R(E'_j, E_{j+1})$ is true and $[\alpha'_{j,u_j}, \dots, \alpha'_{j,j}]$ is type II.
- *Case* 2. $u_j^* = j = u_j > \Gamma + 1$. Now E_j has the form $i_j = \sum_{l=0}^{j-1} \pi_l^* S_{j,l} + \alpha_{j,j}$ and $\lambda_{j,\eta_j} = 0$. In order to have the type II property as before, E_j should be

replaced with a new expression:

$$\widetilde{E}_j: \quad i_j = \sum_{t=0}^{j-2} \pi_t^* S_{j,t} + \widetilde{\alpha}_{j,j-1} S_{j,j-1} + \widetilde{\alpha}_{j,j}, \tag{4.12}$$

where $[\widetilde{\alpha}_{j,j-1}, \widetilde{\alpha}_{j,j}] = D[\pi_{j-1}^*, \alpha_{j,j}]$, i.e.,

$$\widetilde{\alpha}_{j,j-1} = \pi_{j-1}^* - \Delta_j, \ \ \widetilde{\alpha}_{j,j} = \alpha_{j,j} + (q-1)\Delta_j, \ \ \Delta_j = \lfloor (\pi_{j-1}^* - \alpha_{j,j})/q \rfloor.$$

It is easy to see that $[\widetilde{\alpha}_{j,j-1}, \widetilde{\alpha}_{j,j}]$ is type II. However, we do not know if $R(\widetilde{E}_j, E_{j+1})$ is true. In order to make $R(\widetilde{E}_j, E_{j+1})$ true, all of the expressions $E_l (j \le l \le \omega)$ should be changed, where ω is the integer such that

$$j = u_j = u_{j+1} = \cdots = u_\omega > u_{\omega+1}.$$

The new expressions for i_l are given by

$$\widetilde{E}_l: \quad i_l = \sum_{t=0}^{j-2} \pi_t^* S_{l,t} + \sum_{t=j-1}^{l} \widetilde{\alpha}_{l,t} S_{l,t} + \lambda_{l,\eta_l} \quad \text{for} \quad j \le l \le \omega, \tag{4.13}$$

where $\widetilde{\alpha}_{l,j-1} = \pi_{j-1}^* - \Delta_j$ and $\widetilde{\alpha}_{l,t} = \alpha_{l,t} + (q-1)\Delta_j$ for $j \le t \le l$. λ_{l,η_l} is the same as the corresponding term in E_l. Let $\widetilde{u}_l = j-1$. It is easy to verify that $R(\widetilde{E}_l, \widetilde{E}_{l+1})$ is true for $j \le l \le \omega - 1$ and that $R(\widetilde{E}_\omega, E_{\omega+1})$ is also true.

Now, the induction given by Steps 1, 2, and 3 ends the proof. □

Note that, when we construct E_j from E_{j+1} by induction, if Case 2 of Step 3 occurs, then Case 1 of Step 3 will not appear in the next cycle. This is because, in the next cycle, the expression for i_{j-1} obtained by using Step 1 has the form $E_{j-1} : i_{j-1} = \sum_{l=0}^{j-2} \pi_l^* S_{j-1,l} + \alpha_{j-1,j-1}$.

5. Proofs of Theorem 2.4 and two corollaries. The proof of Theorem 2.4 is based on Theorem 2.3 and the following three lemmas: Lemmas 5.1, 5.2, and 5.3. Lemma 5.1 leads to the first part of Theorem 2.4. It tells us how to make use of Theorem 2.3.

LEMMA 5.1. *For a chain permissible sequence* $(a_1, \dots, a_k)$ *and a fixed integer* $\Gamma : 0 \le \Gamma \le k-4$, *if*

$$\pi_\Gamma \ge (k-2)q \quad \text{and} \quad \iota_{k-2} \ge (k-2)(q-1),$$

then there exist integers $\theta_0 \ge \theta_1 \ge \cdots \ge \theta_{k-2} \ge 0$ *such that*

$$i_{k-2} = \sum_{l=0}^{k-2} \theta_l S_{k-2,l}, \quad \text{where} \quad \theta_l \le \pi_l \quad \text{for} \quad 0 \le l \le \Gamma. \tag{5.1}$$

Proof. The proof of Lemma 5.1 is given in two steps. In the first step, an initial expression for i_{k-2} is presented in (5.3). In the second step, the parameters $\theta_0, \dots, \theta_{k-2}$ satisfying (5.1) are obtained in (5.5) and (5.8), respectively. Denote

$$z = \max\left\{ \rho : \ i_{k-2} \ge \sum_{r=0}^{\rho} \pi_r^* S_{k-2,r} \right\} \quad (\text{if } i_{k-2} < \pi_0^* S_{k-2,0}, \text{ let } z = -1), \tag{5.2}$$

where π_r^* is defined in (3.12). If $z = k-2$ or $k-3$, by using Lemma 3.3, the proof is trivial since we can select $\theta_r = \pi_r^*$ for $0 \le r \le k-3$ and $\theta_{k-2} \le \pi_{k-2}^*$. In the following paragraphs, the proof is presented for $z \le k-4$.

First, by using (5.2), an initial expression for i_{k-2} is obtained:

$$i_{k-2} = \sum_{r=0}^{z} \pi_r^* S_{k-2,r} + \sum_{r=z+1}^{k-2} \sigma_r S_{k-2,r}, \tag{5.3}$$

where $\sigma_{z+1}, \dots, \sigma_{k-2}$ are nonnegative integers such that

$$\begin{array}{lll} \sigma_{z+1} & < & \pi_{z+1}^*, \\ \sigma_r & < & S_{k-2,r-1}/S_{k-2,r} = q \quad \text{for} \quad z+2 \le r \le k-3, \\ \sigma_{k-2} & < & S_{k-2,k-3}/S_{k-2,k-2} = q-1. \end{array}$$

In particular, for $z = -1$, we have $i_{k-2} = \sum_{r=0}^{k-2} \sigma_r S_{k-2,r}$, where σ_0 is selected as $\lfloor i_{k-2}/S_{k-2,0} \rfloor$. From the condition $i_{k-2} \ge (k-2)qS_{k-2,0}$, we know that

$$\sigma_0 \ge (k-2)q \quad \text{when} \quad z = -1. \tag{5.4}$$

Second, by adjusting (5.3) in the following two cases, (5.1) is obtained in (5.5) and (5.8), respectively.

- Assume that $\sigma_{z+1} \ge \sigma_{z+2} + (k-z-4)q + 1$; then we have

$$\begin{cases} \theta_r & = \pi_r^* \quad \text{for} \quad 0 \le r \le z, \\ \theta_{z+1} & = \sigma_{z+1} - (k-z-4), \\ \theta_r & = \sigma_r + (k-r-2)q - (k-r-3) \;\; \text{for} \;\; z+2 \le r \le k-3, \\ \theta_{k-2} & = \sigma_{k-2}. \end{cases} \tag{5.5}$$

  This assumption implies that $\theta_{z+1} \ge \theta_{z+2}$. In addition, the condition $(k-2)q \le \pi_\Gamma$ implies that $\theta_{z+2} \le \pi_\Gamma$ since $\theta_{z+2} \le (k-2)q$.

- Assume that $\sigma_{z+1} < \sigma_{z+2} + (k-z-4)q + 1$; we have

$$z \ge 0. \tag{5.6}$$

  If $z = -1$, then the assumption denotes that $\sigma_0 < \sigma_1 + (k-3)q + 1 \le (k-2)q$, which is opposite to (5.4). Let $\mu = \lfloor \sigma_{z+1}/q \rfloor$; then

$$k - z - \mu - 3 \ge 0. \tag{5.7}$$

  If $k-z-\mu-3 < 0$, then the assumption denotes that $\sigma_{z+1} < (k-z-3)q \le \mu q$, which is also impossible. By using (5.6), (5.7), and the condition $(k-2)q \le \pi_\Gamma$, we have

$$\begin{cases} \theta_r & = \pi_r^* \quad \text{for} \quad 0 \le r \le z-1, \\ \theta_z & = \pi_z^* - (k-z-\mu-3), \\ \theta_{z+1} & = \sigma_{z+1} + (k-z-\mu-3)q - (k-z-4), \\ \theta_r & = \sigma_r + (k-r-2)q - (k-r-3) \;\; \text{for} \;\; z+2 \le r \le k-3, \\ \theta_{k-2} & = \sigma_{k-2}. \end{cases} \tag{5.8}$$

  Note that, in (5.8), since $\theta_{z+1} = (\sigma_{z+1} - \mu q) + (k-z-3)q - (k-z-4)$, we have

$$(k-z-2)q - (k-z-3) \ge \theta_{z+1} \ge (k-z-3)q - (k-z-4),$$

  which implies that $\theta_z \ge \theta_{z+1} \ge \theta_{z+2}$ and $\pi_\Gamma \ge \theta_{z+1}$. $\square$

In the following lemma, i.e., Lemma 5 of [1], a relation between the parameters ι_l and π_l is introduced. Using this lemma, the second part of Theorem 2.4 can be obtained from the first part of Theorem 2.4.

LEMMA 5.2 (see [1]). *For a chain permissible sequence with dimension k, if there exists a positive integer l such that $\iota_r = \iota_{r+1} + \delta_r$ for $0 \le r \le l-1$, then*

$$\iota_l = \pi_l + \sum_{r=0}^{l-1} (\delta_r(q^{r+1} - 1) + qp_r - p_{r+1}). \tag{5.9}$$

Lemma 5.3, a special case of Lemma 8 of [6], allows us to pay attention to some special chain permissible sequences.

LEMMA 5.3. *For fixed nonnegative integers $l(\le k-1)$, s, and F, if each chain permissible sequence such that*

$$\iota_j = \iota_{j+1} + \delta_j \quad \textit{for} \quad 0 \le j \le l-1, \tag{5.10}$$

$$\iota_l \ge s + \sum_{r=0}^{l-1} (\delta_r(q^{r+1} - 1) + qp_r - p_{r+1}), \tag{5.11}$$

$$i_{j-1} \ge i_j/q + S_{j-1,0} \quad \textit{for} \quad l+2 \le j \le k-3, \tag{5.12}$$

$$i_{k-2} \ge F \tag{5.13}$$

is a chain good weight hierarchy, then the chain permissible sequences which satisfy only (5.11), (5.12), *and* (5.13) *are chain good weight hierarchies.*

Proof of Theorem 2.4. For $0 \le \Gamma \le k-4$, by Lemma 5.1 and Theorem 2.3, we know that the chain permissible sequences such that

$$\pi_\Gamma \ge (k-2)q, \quad \iota_{k-2} \ge (k-2)(q-1),$$

and

$$i_{j-1} \ge i_j/q + S_{j-1,0} \quad \text{for} \quad \Gamma + 2 \le j \le k-3 \tag{5.14}$$

are chain good weight hierarchies. This is the first part of Theorem 2.4. Then by Lemma 5.2, the chain permissible sequences such that

$$\begin{aligned} &\iota_u = \iota_{u+1} + \delta_u \quad \text{for} \quad 0 \le u \le \Gamma - 1,\\ &\iota_\Gamma \ge (k-2)q + \sum_{r=0}^{\Gamma-1} (\delta_r(q^{r+1} - 1) + qp_r - p_{r+1}),\\ &i_{k-2} \ge (k-2)qS_{k-2,0},\\ &i_{j-1} \ge i_j/q + S_{j-1,0} \quad \text{for} \quad \Gamma + 2 \le j \le k-3 \end{aligned} \tag{5.15}$$

are chain good weight hierarchies. Finally, by using Lemma 5.3 with parameters $l = \Gamma$, $s = (k-2)q$, and $F = (k-2)qS_{k-2,0}$, the second part of this theorem is obtained. Note that, for $\Gamma = k-4$, conditions (5.14) and (5.15) do not exist. □

Proof of Corollary 2.5. Corollary 2.5 follows from the second part of Theorem 2.4. Condition (2.15) is satisfied by (2.17). Condition (2.16) can be obtained by using (2.18) and the inequality $i_{j-1}/q^{j-1} \ge \iota_{j-1} \ge \iota_j + 2 \ge i_j/q^j + 1$, where $\Gamma + 2 \le j \le k-3$. We will show that the condition (2.14) is also satisfied.

For a chain permissible sequence $(a_1, \ldots, a_k)$, it follows from (2.7) and (2.18) that

$$\iota_\Gamma \ge \iota_{\Gamma+1} \ge \iota_{k-3} + 2(k-4-\Gamma) \ge \iota_{k-2} + 2(k-4-\Gamma). \tag{5.16}$$

Then by using (5.16) and (2.17), we have

$$\iota_\Gamma \geq (k-2)(q-1) + \sum_{r=0}^{\Gamma-1} q^{r+1} + 2(k-4-\Gamma) \geq (k-2)q + \sum_{r=0}^{\Gamma-1}(q^{r+1}-2)$$

since $k \geq \Gamma + 6$. Therefore, (2.14) is satisfied since

$$q^{r+1} - 2 \geq \delta_r(q^{r+1}-1) + qp_r - p_{r+1}. \qquad \square$$

The proof of Corollary 2.6 uses the same arguments as that of Corollary 2.5.

6. Improvements on [1] and [6]. Theorem 2.4 presents a series of sufficient conditions for determining the chain good weight hierarchies by using different Γ's. In this section, using Theorem 2.4, we find many new chain good weight hierarchies, which cannot be investigated using Theorems 2.1 and 2.2. For $q = 3$ and $k = 6, 7, 8$, three examples of the improvements are given by using Corollaries 6.1, 6.2, and 6.3, respectively.

Let $(a_1, \ldots, a_k)$ be a chain permissible sequence and let Γ be an integer such that $0 \leq \Gamma \leq k-4$. From the second part of Theorem 2.4, we know that $(a_1, \ldots, a_k)$ is chain good if (2.14), (2.15), and (2.16) are satisfied. Since $\delta_r(q^{r+1}-1)+qp_r-p_{r+1} \leq q^{r+1}-2$, it is easy to see that a chain permissible $(a_1, \ldots, a_k)$ is chain good if

$$\begin{aligned} &\iota_\Gamma \geq (k-2)q + \sum_{r=0}^{\Gamma-1}(q^{r+1}-2), \\ &\iota_{k-2} \geq (k-2)(q-1), \end{aligned}$$

and

$$i_{j-1} \geq i_j/q + S_{j-1,0} \quad \text{for} \quad \Gamma + 2 \leq j \leq k-3.$$

Then, Corollaries 6.1, 6.2, and 6.3 are obtained for $q = 3$ and $\Gamma = k - 5 = 1$, $\Gamma = k - 6 = 1$, and $\Gamma = k - 7 = 1$, respectively.

Corollary 6.1. *For $q = 3$ and $k = 6$, a chain permissible sequence is a chain good weight hierarchy if*

$$\iota_1 \geq 13, \quad \iota_4 \geq 8, \quad \textit{and} \quad i_2 \geq i_3/3 + 6. \tag{6.1}$$

Example. From Corollary 6.1, we find that, for each pair of parameters (i_3, i_4) such that $648 \leq i_4 \leq 1997$ and $i_4/3 \leq i_3 \leq 695$, there exist many new chain good weight hierarchies which cannot be investigated using Theorems 2.1 and 2.2. For instance, if $i_4 = 648$ and $i_3 = 216$, all the corresponding chain permissible sequences with dimension 6 such that $i_2 \in \{115, 116\} \bigcup \{l : 120 + 9t \leq l \leq 125 + 9t, 0 \leq t \leq 13\}$ are new chain good weight hierarchies.

Corollary 6.2. *For $q = 3$ and $k = 7$, a chain permissible sequence is a chain good weight hierarchy if*

$$\iota_1 \geq 16, \quad \iota_5 \geq 10, \quad \textit{and} \quad i_{j-1} \geq i_j/3 + 2 \cdot 3^{j-2} \textit{ for } j = 3, 4. \tag{6.2}$$

Example. From Corollary 6.2, we find that, for each pair of parameters (i_4, i_5) satisfying $2430 \leq i_5 \leq 19013$ and $i_5/3 \leq i_4 \leq 6419$, many new chain good weight hierarchies cannot be checked with Theorems 2.1 and 2.2. For instance, if $i_5 = 2430$

and $i_4 = 810$, all the corresponding chain permissible sequences with dimension 7 such that $i_2 \geq i_3/3 + 6$ and $i_3 \in \{l : 409 + 27t \leq l \leq 431 + 27t, 0 \leq t \leq 65\}$ are new chain good weight hierarchies.

COROLLARY 6.3. *For $q = 3$ and $k = 8$, a chain permissible sequence is a chain good weight hierarchy if*

$$\iota_1 \geq 19, \quad \iota_6 \geq 12, \quad \textit{and} \quad i_{j-1} \geq i_j/3 + 2 \cdot 3^{j-2} \textit{ for } j = 3, 4, 5. \tag{6.3}$$

Example. From Corollary 6.3, we find that, for each pair of parameters (i_5, i_6) such that $8748 \leq i_6 \leq 174695$ and $i_6/3 \leq i_5 \leq 58475$, there are also many new chain good weight hierarchies which cannot be investigated using Theorems 2.1 and 2.2. For instance, if $i_6 = 8748$ and $i_5 = 2916$, all the corresponding chain permissible sequences with dimension 8 such that $i_2 \geq i_3/3 + 6$, $i_3 \geq i_4/3 + 18$, and $i_4 \in \{l : 1405 \leq l \leq 1457\} \bigcup \{l : 1463 + 81t \leq l \leq 1538 + 81t, 0 \leq t \leq 224\}$ are new chain good weight hierarchies.

7. Conclusion. The determination of chain good weight hierarchies was studied several years ago. For the binary codes with dimension up to 5 and the ternary codes with dimension up to 4, the problem was solved in [3] and [2], respectively. As for linear codes with general dimension over $GF(q)$, some research was done in [1] and [6]. However, these results are not efficient for the determination of the chain good weight hierarchies with high dimension since in many cases the lower bounds on the conditions for $\iota_0, \ldots, \iota_{k-3}(\text{or } \iota_{k-2})$ increase exponentially with the dimension k. In this paper, we present a method to deal with the high dimension cases; see Corollaries 2.5 and 2.6. Our lower bounds on the conditions for $\iota_0, \ldots, \iota_{k-2}$ only increase linearly with the dimension k.

REFERENCES

[1] W. CHEN AND T. KLØVE, *Weight hierarchies of linear codes satisfying the chain condition*, Des. Codes Cryptogr., 15 (1998), pp. 47–66.

[2] W. CHEN AND T. KLØVE, *The weight hierarchies of q-ary codes of dimension* 4, IEEE Trans. Inform. Theory, 42 (1996), pp. 2265–2272.

[3] S. ENCHEVA AND T. KLØVE, *Codes satisfying the chain condition*, IEEE Trans. Inform. Theory, 40 (1994), pp. 175–180.

[4] T. HELLESETH, T. KLØVE, AND J. MYKKELTVEIT, *The weight distribution of irreducible cyclic codes with block lengths* $n_1((q^l - 1)/N)$, Discrete Math., 18 (1977), pp. 179–211.

[5] T. HELLESETH, T. KLØVE, AND Ø. YTREHUS, *Generalized Hamming weights of linear codes*, IEEE Trans. Inform. Theory, 38 (1992), pp. 1133–1140.

[6] Y. LUO, W. CHEN, AND F. FU, *A new kind of geometric structures determining the chain good weight hierarchies*, Discrete Math., 260 (2003), pp. 101–117.

[7] V. K. WEI, *Generalized Hamming weight for linear codes*, IEEE Trans. Inform. Theory, 37 (1991), pp. 1412–1418.

[8] V. K. WEI AND K. YANG, *On the generalized Hamming weight of product codes*, IEEE Trans. Inform. Theory, 39 (1993), pp. 1709–1713.

SIAM J. Discrete Math.
Vol. 17, No. 2, pp. 210–218

SOLUTIONS FOR TWO CONJECTURES ON THE INVERSE PROBLEM OF THE WIENER INDEX OF PEPTOIDS[*]

XUELIANG LI[†] AND LUSHENG WANG[‡]

Abstract. In this paper, we give solutions for the two conjectures on the inverse problem of the Wiener index of peptoids proposed by Goldman et al. We give the first conjecture a positive proof and the second conjecture a negative answer.

Key words. combinatorial chemistry, Wiener index, peptoid

AMS subject classification. 92E10

DOI. 10.1137/S0895480101387261

1. Introduction.

In drug design and molecular recognition, combinatorial chemistry has played a powerful role in recent years. One of the central problems is the construction of a molecular graph with given chemical or physical properties. A chemical or physical property can be quantitatively represented by some topological index [1]. The problem here is to find a molecular graph with a given value of some topological index. In [1], the authors studied the problem for the Wiener index. They proposed two conjectures related to the so-called *inverse problem of peptoids*. A peptoid is represented by a large molecular graph constructed from some pieces of given small molecular graphs by joining them in a linear scaffold way, i.e., chaining them linearly. The problem is to find a peptoid with these given small pieces as fragments such that it has the desired Wiener index value. The ordering or arrangement of these pieces in a peptoid determines the value of the Wiener index. The two conjectures are to determine the orderings or arrangements under which the values are minimum or maximum. As one can see in the statements of the conjectures, the optimal problems are purely mathematical. We can go without any notation or terminology on graph theory or chemistry.

Let $n_1, n_2, \ldots, n_N$ be N positive integers; define

$$D = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (j-i) n_i n_j.$$

For an ordering or rearrangement $\pi = \pi(1)\pi(2)\ldots\pi(N)$ of $1, 2, \ldots, N$, define

$$D(\pi) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (j-i) n_{\pi(i)} n_{\pi(j)}.$$

Conjecture 5.1 of [1] is stated as follows.

[*] Received by the editors April 3, 2001; accepted for publication (in revised form) May 28, 2003; published electronically November 4, 2003.
http://www.siam.org/journals/sidma/17-2/38726.html

[†] Department of Applied Mathematics, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China and The Key Laboratory of the Chinese Education Ministry for Pure Mathematics and Combinatorics, Research Center for Combinatorics, Nankai University, Tianjin 300071, China. The research of this author was partially supported by the Natural Science Foundation of China.

[‡] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (lwang@cs.cityu.edu.hk). The research of this author was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, project CityU 1087/00E.

CONJECTURE 1. *Given $n_1 \leq n_2 \leq \cdots \leq n_N$, the ordering for the minimum value of D is*

$$\pi_{min}(i) = \begin{cases} 2i-1 & \textit{if } i \leq \frac{N}{2} \\ 2(N-i+1) & \textit{if } i > \frac{N}{2}. \end{cases}$$

Conjecture 5.2 of [1] is stated as follows.

CONJECTURE 2. *An algorithm to compute the ordering for the maximum value of D, given $n_1 \leq n_2 \leq \cdots \leq n_N$, is as follows:*

```
$L_P = 1; L = 0$
$R_p = N; R = 0$
For $i = N$ down to $1$ do
  if $R \geq L$, then
   $\pi_{max}(L_p) = i; L_p = L_p + 1; L = L + n_i;$
  else
   $\pi_{max}(R_p) = i; R_p = R_p - 1; R = R + n_i.$
```

We solve these two conjectures in the following sections. In section 2, we do some preparations by introducing two inequalities due to Hardy, Littlewood, and Pólya [2] and Wiener [4], respectively. In section 3, we prove Conjecture 1 vigorously by using Hardy, Littlewood, and Pólya's inequality. In section 4, we show that Conjecture 2 is not correct. The algorithm in Conjecture 2 does not always give the maximum value. We give a better upper bound for the maximum value by using Wiener's inequality. Finally, in section 5, we analyze the difficulty in finding the exact ordering to attain the maximum value.

2. Preliminaries. We follow the notations of [2] or [3]. Suppose that we are given a set of a finite number of nonnegative numbers $x_1, x_2, \ldots, x_N$, or $x_{-n}, \ldots, x_{-1}, x_0, x_1, \ldots, x_n$, denoted by (x). An ordering or rearrangement of them is $x'_1, x'_2, \ldots, x'_N$, or $x'_{-n}, \ldots, x'_{-1}, x'_0, x'_1, \ldots, x'_n$, denoted by (x'), where $\{x'_1, x'_2, \ldots, x'_N\} = \{x_1, x_2, \ldots, x_N\}$ and $\{x'_{-n}, \ldots, x'_{-1}, x'_0, x'_1, \ldots, x'_n\} = \{x_{-n}, \ldots, x_{-1}, x_0, x_1, \ldots, x_n\}$. Some special orderings are given as follows:

$$(\bar{x}) = \bar{x}_1 \leq \bar{x}_2 \leq \cdots \leq \bar{x}_N$$

or

$$(\bar{x}) = \bar{x}_{-n} \leq \cdots \leq \bar{x}_{-1} \leq \bar{x}_0 \leq \bar{x}_1 \leq \cdots \leq \bar{x}_n,$$

i.e., increasing ordering.

$$(x^+) = x_0^+ \geq x_1^+ \geq x_{-1}^+ \geq x_2^+ \geq x_{-2}^+ \geq \cdots$$

and

$$({}^+x) =^+ x_0 \geq^+ x_{-1} \geq^+ x_1 \geq^+ x_{-2} \geq^+ x_2 \geq \cdots.$$

For example, in the example of [1, p. 283], $(x) = 8, 13, 2, 17, 19, 18, 28, 5$; $(\bar{x}) = (n) = 2, 5, 8, 13, 17, 18, 19, 28$; $(x^+) = 5, 13, 18, 28, 19, 17, 8, 2$, and $({}^+x) = 2, 8, 17, 19, 28, 18, 13, 5$.

From [2] or [3], we have the following theorem.

Theorem 2.1 (Hardy, Littlewood and Pólya). *Suppose that c, x, y are nonnegative and c symmetrically decreasing so that*

$$c_0 \geq c_1 = c_{-1} \geq c_2 = c_{-2} \geq \cdots \geq c_{2k} = c_{-2k},$$

while x and y are given except in arrangement. Then the bilinear form

$$S^{(1)} = \sum_{r=-k}^{k} \sum_{s=-k}^{k} c_{r-s} x_r y_s \tag{1}$$

attains its maximum when (x) is (x^+) and (y) is (y^+), or (x) is (^+x) and (y) is (^+y).

From [2] or [4], we have the following theorem.

Theorem 2.2 (Wiener). *If $c_2 \geq c_3 \geq \cdots \geq c_{2n} \geq 0$ and the sets (x) and (y) are nonnegative and given except in arrangement, then*

$$S^{(2)} = \sum_{r=1}^{n} \sum_{s=1}^{n} c_{r+s} x_r y_s \tag{2}$$

is a maximum when (x) and (y) are both in decreasing order.

It is easy to see that the two bilinear forms of (1) and (2) have the coefficient matrices

$$C^{(1)} = \begin{pmatrix} c_0 & c_1 & c_2 & c_3 & \cdots & \cdots & \cdots & \cdots \\ c_1 & c_0 & c_1 & c_2 & \cdots & \cdots & \cdots & \cdots \\ c_2 & c_1 & c_0 & c_1 & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & c_0 & c_1 & c_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & c_1 & c_0 & c_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & c_2 & c_1 & c_0 \end{pmatrix}_{(2k+1)\times(2k+1)}$$

and

$$C^{(2)} = \begin{pmatrix} c_2 & c_3 & c_4 & c_5 & \cdots & \cdots & \cdots & c_{n+1} \\ c_3 & c_4 & c_5 & \cdots & \cdots & \cdots & \cdots & \cdots \\ c_4 & c_5 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ c_5 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & c_{2n-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & c_{2n-2} & c_{2n-1} \\ c_{n+1} & \cdots & \cdots & \cdots & \cdots & c_{2n-2} & c_{2n-1} & c_{2n} \end{pmatrix}_{n\times n},$$

respectively.

So, we have

$$S^{(1)} = \ (x_{-k}, \ldots, x_{-1}, x_0, x_1, \ldots, x_k) \ C^{(1)} \begin{pmatrix} y_{-k} \\ \vdots \\ y_{-1} \\ y_0 \\ y_1 \\ \vdots \\ y_k \end{pmatrix}$$

and

$$S^{(2)} = \ (x_1, x_2, \dots, x_n) \ C^{(2)} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

3. The proof of Conjecture 1. We shall use Theorem 2.1 to prove Conjecture 1. Since Conjecture 1 is about optimal minimum, while Theorem 2.1 is about optimal maximum, we have to do some transformation in the following.

First, we note that

$$2D = \sum_{i=1}^{n} \sum_{j=1}^{n} |j-i| n_i n_j$$

with the coefficient matrix as follows:

$$A = \begin{pmatrix} 0 & 1 & 2 & \cdots & N-2 & N-1 \\ 1 & 0 & 1 & \cdots & N-3 & N-2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ N-2 & N-3 & N-4 & \cdots & 0 & 1 \\ N-1 & N-2 & N-3 & \cdots & 1 & 0 \end{pmatrix}_{N \times N.}$$

Then,

$$D = \frac{1}{2} \ (n_1, n_2, \dots, n_N) \ A \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_N \end{pmatrix}.$$

Take the matrix $C^{(1)} = NI_N - A$, where I_N is the identity of order N, and consider the following bilinear (quadratic) form:

$$\begin{aligned} & (n_1, n_2, \dots, n_N) \ C^{(1)} \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_N \end{pmatrix} \\ &= \ (n_1, n_2, \dots, n_N) \ NI_N \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_N \end{pmatrix} - \ (n_1, n_2, \dots, n_N) \ A \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_N \end{pmatrix} \\ &= N \left(\sum_{i=1}^{N} n_i \right)^2 - 2D. \qquad (3) \end{aligned}$$

Since the term $N(\sum_{i=1}^{n} n_i)^2$ in (3) is independent of the orderings of $n_1, n_2, \dots, n_N$, we have that (3) reaches its optimal maximum by some ordering of $n_1, n_2, \dots, n_N$ if and only if D reaches its optimal minimum. Since here the matrix $C^{(1)}$ satisfies the conditions of Theorem 2.1, we know that (3) reaches its optimal maximum when (n) is (n^+) or (^+n), which is exactly the ordering given in Conjecture 1. Therefore, D reaches its optimal minimum when (n) is the ordering given in Conjecture 1. The proof is complete. □

4. Negative answer for Conjecture 2 and a better upper bound. We use examples to give a negative answer for Conjecture 2. The example of [1, p. 283] is shown in Table 1.

Table 1
The table of [1, *p.* 283].

i	1	2	3	4	5	6	7	8
n_i	2	5	8	13	17	18	19	28
$n_{\pi_{max}(i)}$	28	**18**	8	2	5	13	**17**	19

First, we point out that by executing the algorithm in Conjecture 2, we get a different ordering from Table 1, with the exchange of the two numerals 18 and 17 in the third line. We denote the ordering in Table 1 by π_t, not by π_{max}, and the ordering determined by the algorithm of Conjecture 2 by π_a. So $\pi_t = 8, 6, 3, 1, 2, 4, 5, 7$. In Appendix A, we show that $\pi_a = 8, 5, 3, 1, 2, 4, 6, 7$ by executing the algorithm. We do not know if this π_a gives the optimal maximum. However, we do know that $D(\pi_a) > D(\pi_t)$. In fact, we have that

$$\begin{aligned} D(\pi_t) - D(\pi_a) &= (n_6 - n_5)[5(n_7 - n_8) + 3(n_4 - n_3) + (n_2 - n_1)] \\ &= (18 - 17)[5(19 - 28) + 3(13 - 8) + (5 - 2)] \\ &= 5 \times (-9) + 3 \times 5 + 3 \\ &= -45 + 18 = -27 < 0, \end{aligned} \tag{4}$$

i.e., $D(\pi_a) > D(\pi_t)$.

Does this mean that the algorithm in Conjecture 2 really gives the ordering for the optimal maximum? The answer is “no.” One may argue that the ordering of the numerals given in the table of [1, p. 283] is misprinted by the authors’ carelessness. This is also not the case. In fact, from (4) we can see that $n_6 - n_5$ is always positive when $n_5 \neq n_6$, and so are $n_4 - n_3$ and $n_2 - n_1$. However, $n_7 - n_8$ is always negative when $n_7 \neq n_8$. One can imagine that by properly assigning the values of $n_1, n_2, \ldots, n_8$, we can get $D(\pi_a) > D(\pi_t)$, as in the above example, and $D(\pi_a) < D(\pi_t)$ in some other cases. This is really the case. For example, we take $n_1 = 1$, or any number smaller than 8, $n_2 = 20$, $n_3 = 21, n_4 = 22, n_5 = 23, n_6 = 24, n_7 = 25, n_8 = 28$. The ordering given by π_t and the ordering π_a obtained by the algorithm of Conjecture 2 are shown in Table 2.

Table 2
The orderings given by π_t *and* π_a*, respectively.*

i	1	2	3	4	5	6	7	8
n_i	1	20	21	22	23	24	25	28
$n_{\pi_t(i)}$	28	**24**	21	1	20	22	**23**	25
$n_{\pi_a(i)}$	28	**23**	21	1	20	22	**24**	25

From (4), we have

$$\begin{aligned} D(\pi_t) - D(\pi_a) &= (24 - 23)[5(25 - 28) + 3(22 - 21) + (20 - 1)] \\ &= 5 \times (-3) + 3 \times 1 + 19 = 7 > 0, \end{aligned}$$

i.e., $D(\pi_t) > D(\pi_a)$. In fact, since $n_6 - n_5 > 0$, from (4) we know that $D(\pi_t) > D(\pi_a)$ if $5(n_8 - n_7) < 3(n_4 - n_3) + (n_2 - n_1)$, while $D(\pi_t) < D(\pi_a)$ if $5(n_8 - n_7) > 3(n_4 - n_3) + (n_2 - n_1)$. So we can construct infinitely many examples to show that $D(\pi_t) > D(\pi_a)$ by properly assigning the values of $n_1, n_2, \ldots, n_8$ and also infinitely many other examples to show that $D(\pi_t) < D(\pi_a)$, the other way round. We do not know that if this $D(\pi_t)$ is the optimal maximum for these new n_i's. However, it is greater than the value under the ordering given by the algorithm of Conjecture 2.

So the ordering with optimal maximum value is still unknown. We tried to find it but failed. First, we look at the ordering given by Conjecture 1, which attains the optimal minimum. Imagine that we have a balance, or a rod with support point at the center, and we want to hang N things with weights $n_1, n_2, \ldots, n_N$ on it. To reach the minimum, we hang the heaviest one n_N at the (near) center, then we take turns to hang the next heaviest things to the left (or right) and right (or left) side of n_N. One may imagine that the maximum might be attained the other way round, i.e., hang the lightest one n_1 at the (near) center, then take turns to hang the next lightest things to the left (or right) and right (or left) side of n_1. We denote this ordering by π_b. Unfortunately, for $N = 8$ we have that

$$D(\pi_a) - D(\pi_b) = (n_8 - n_7)[5(n_6 - n_5) + 3(n_4 - n_3) + (n_2 - n_1)] > 0$$

and

$$D(\pi_t) - D(\pi_b) = [(n_8 - n_7) + (n_6 - n_5)][3(n_4 - n_3) + (n_2 - n_1)] > 0.$$

i.e.,

$$D(\pi_b) < D(\pi_a) \quad \text{and} \quad D(\pi_b) < D(\pi_t).$$

So the intuitive observation does not give an ordering with the optimal maximum. Indeed, finding such an ordering is not an easy thing. This is why the authors of [1] did not give an exact ordering but instead an algorithmic ordering. The above analysis shows us that, unlike the optimal minimum case, to obtain the optimal maximum, the ordering is not purely dependent on the value-ordering of $n_1, n_2, \ldots, n_N$ but mainly dependent on how large the values $n_1, n_2, \ldots, n_N$ are themselves. Although to give an exact ordering for the optimal maximum is almost hopeless (see the analysis in section 5), we can give a better upper bound for the optimal maximum by Theorem 2.2, which could be much better than the upper bound $\frac{N^3-N}{6} n_N^2$ in [1] and useful in the inverse problem for peptoids. First, we note that we can rewrite D as follows:

$$D = (n_1, n_2, \ldots, n_N) \begin{pmatrix} N-1 & N-2 & N-3 & \cdots & 2 & 1 & 0 \\ N-2 & N-3 & N-4 & \cdots & 1 & 0 & \\ \vdots & \vdots & \vdots & \vdots & & & \\ 2 & 1 & 0 & & & & \\ 1 & 0 & & & & & \\ 0 & & & & & & \end{pmatrix} \begin{pmatrix} n_N \\ n_{N-1} \\ \vdots \\ n_2 \\ n_1 \end{pmatrix}.$$

Denote the coefficient matrix by $C^{(2)}$. Then, $C^{(2)}$ satisfies the conditions of

Theorem 2.2. Therefore, for any ordering π of $1, 2, \cdots, N$, we have

$$D(\pi) \leq (n_N, n_{N-1}, \ldots, n_2, n_1) C^{(2)} \begin{pmatrix} n_N \\ n_{N-1} \\ \vdots \\ n_2 \\ n_1 \end{pmatrix}.$$

So, we have proved the following theorem.

THEOREM 4.1.

$$D_{max} \leq (n_N, n_{N-1}, \ldots, n_2, n_1) \cdot \begin{pmatrix} N-1 & N-2 & N-3 & \cdots & 2 & 1 & 0 \\ N-2 & N-3 & N-4 & \cdots & 1 & 0 & \\ \vdots & \vdots & \vdots & \vdots & & & \\ 2 & 1 & 0 & & & & \\ 1 & 0 & & & & & \\ 0 & & & & & & \end{pmatrix} \begin{pmatrix} n_N \\ n_{N-1} \\ \vdots \\ n_2 \\ n_1 \end{pmatrix}.$$

5. Difficulty analysis for finding the ordering π_{max}. From the three orderings π_t, π_a, and π_b, we observed that all of them arrange the values $n_1, n_2, \ldots, n_N$ concavely with the valley at the (nearly) central position. Can it be true that the optimal maximum is always attained by some ordering that arranges $n_1, n_2, \ldots, n_N$ in a concave way? The following analysis shows in some extent that the answer is "no." This negative answer shows that in some sense finding the ordering π_{max} could be very difficult.

Suppose that we have a concave ordering π for n_1, n_2, $\ldots$, n_N such that $\pi(i) > \pi(j)$ when $i < j \leq \lfloor \frac{N}{2} \rfloor + 1$, where $\lfloor x \rfloor$ denotes the maximum integer less than or equal to x. We construct another ordering π' from π by

$$\pi'(k) = \begin{cases} \pi(k), & k \neq i, j, \\ \pi(j), & k = i, \\ \pi(i), & k = j, \end{cases}$$

i.e., by exchanging $\pi(i)$ and $\pi(j)$ and keeping the others unchanged. Then, π' is no longer a concave ordering for $n_1, n_2, \ldots, n_N$. We shall show that sometimes $D(\pi) > D(\pi')$ and sometimes $D(\pi) < D(\pi')$, the other way round. First, by careful calculation, we can obtain that

$$\begin{aligned} D(\pi') - D(\pi) &= \sum_{k=1}^{i-1} (j-i) n_{\pi(k)} (n_{\pi(i)} - n_{\pi(j)}) \\ &\quad + \sum_{k=i+1}^{j-1} (2k - (i+j)) n_{\pi(k)} (n_{\pi(j)} - n_{\pi(i)}) \\ &\quad + \sum_{k=j+1}^{N} (j-i) n_{\pi(k)} (n_{\pi(j)} - n_{\pi(i)}) \\ &= (n_{\pi(j)} - n_{\pi(i)}) \sum_{k=1, k \neq i,j}^{N} \alpha_k n_{\pi(k)}, \end{aligned}$$

where

$$\alpha_k = \begin{cases} j-i, & k=1,2,\dots,i-1, \\ (i+j)-2k, & k=i+1,\dots,j-1, \\ -(j-i), & k=j+1,\dots,N. \end{cases}$$

Note that $n_{\pi(i)} - n_{\pi(j)} > 0$ by our assumption that $n_{\pi(i)} > n_{\pi(j)}$.

Example 5.1. When $j = i+1$, we have

$$\frac{D(\pi') - D(\pi)}{n_{\pi(i)} - n_{\pi(i+1)}} = n_{(\pi(1)} + n_{\pi(2)} + \cdots + n_{\pi(i-1)}) - (n_{\pi(i+2)} + n_{\pi(i+3)} + \cdots + n_{\pi(N)}).$$

Example 5.2. When $j = i+2$, we have

$$\begin{aligned}\frac{D(\pi') - D(\pi)}{n_{\pi(i)} - n_{\pi(i+2)}} &= 2n_{\pi(1)} + 2n_{\pi(2)} + \cdots + 2n_{\pi(i-1)} - 2n_{\pi(i+3)} - 2n_{\pi(i+4)} - \cdots - 2n_{\pi(N)} \\ &= 2[(n_{\pi(1)} + n_{\pi(2)} + \cdots + n_{\pi(i-1)}) - (n_{\pi(i+3)} + n_{\pi(i+4)} + \cdots + n_{\pi(N)})].\end{aligned}$$

Example 5.3. When $j = i+3$, we have

$$\begin{aligned}\frac{D(\pi') - D(\pi)}{n_{\pi(i)} - n_{\pi(i+3)}} &= 3n_{\pi(1)} + 3n_{\pi(2)} + \cdots + 3n_{\pi(i-1)} + n_{\pi(i+1)} \\ &\quad - n_{\pi(i+2)} - 3n_{\pi(i+4)} - 3n_{\pi(i+5)} - \cdots - 3n_{\pi(N)} \\ &= 3[(n_{\pi(1)} + n_{\pi(2)} + \cdots + n_{\pi(i-1)}) - (n_{\pi(i+4)} + n_{\pi(i+5)} + \cdots + n_{\pi(N)})] \\ &\quad + (n_{\pi(i+1)} - n_{\pi(i+2)}).\end{aligned}$$

From Examples 5.1–5.3, we can see that one can properly assign the values $n_1, n_2, \dots, n_N$ to attain $D(\pi') < D(\pi)$ sometimes, or $D(\pi') > D(\pi)$ on other occasions. This again shows that the most important aspect for the ordering to attain the optimal maximum is heavily dependent on how large the value is itself of each of the $n_1, n_2, \dots, n_N$, and is not purely dependent on the value-ordering of them. In other words, different values of $n_1 \leq n_2 \leq \cdots \leq n_N$ give different orderings for attaining the optimal maximum.

To conclude the paper we propose the following problem.

Problem 5.1. Find a polynomial-time algorithm to compute the ordering for the optimal maximum of D, given $n_1 \leq n_2 \leq \cdots \leq n_N$.

Appendix A. We follow the algorithm of Conjecture 2 for the numerals n_i in Table 1.

Step 0. $L_p = 1, L = 0; R_p = 8, R = 0$

Step 1. $i = 8; R = 0 \geq 0 = L,$
$\pi_{max}(L_p) = \pi_{max}(1) = 8;$
$L_p = 1 + 1 = 2, L = 0 + n_8 = 28$

Step 2. $i = 7; R = 0 < L = n_8 = 28,$
$\pi_{max}(R_p) = \pi_{max}(8) = 7;$
$R_p = 8 - 1 = 7, R = 0 + n_7 = 19$

Step 3. $i = 6; R = 19 < 28 = L,$

$\pi_{max}(R_p) = \pi_{max}(7) = 6;$
$R_p = 7 - 1 = 6, R = 19 + n_6 = 19 + 18 = 37$

Step 4. $i = 5; R = 37 > 28 = L,$
$\pi_{max}(L_p) = \pi_{max}(2) = 5;$
$L_p = 2 + 1 = 3, L = 28 + n_5 = 28 + 17 = 45$

Step 5. $i = 4; R = 37 < 45 = L,$
$\pi_{max}(R_p) = \pi_{max}(6) = 4;$
$R_p = 6 - 1 = 5, R = 37 + n_4 = 37 + 13 = 50$

Step 6. $i = 3; R = 50 > 45 = L,$
$\pi_{max}(L_p) = \pi_{max}(3) = 3;$
$L_p = 3 + 1 = 4, L = 45 + n_3 = 45 + 8 = 53$

Step 7. $i = 2; R = 50 < 53 = L,$
$\pi_{max}(R_p) = \pi_{max}(5) = 2;$
$R_p = 5 - 1 = 4, R = 50 + n_2 = 50 + 5 = 55$

Step 8. $i = 1; R = 55 > 53 = L,$
$\pi_{max}(L_p) = \pi_{max}(4) = 1;$
$L_p = 4 + 1 = 5, L = 53 + n_1 = 53 + 2 = 55$

Finally, we get an ordering π_{max} or $\pi_a = 8, 5, 3, 1, 2, 4, 6, 7$.

REFERENCES

[1] D. Goldman, S. Istrail, G. Lancia, A. Piccolboni, and B. Walenz, *Algorithmic strategies in combinatorial chemistry*, in Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 2000, pp. 275–284.

[2] G. Hardy, J.E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed., Cambridge University Press, Cambridge, UK, 1988, pp. 261–265, p. 295.

[3] G. Hardy, J.E. Littlewood, and G. Pólya, *The maximum of a certain bilinear form*, Proc. London Math. Soc. (2), 25 (1926), pp. 265–282.

[4] F. Wiener, *Elementarer beweis eines reihensatzes von Herrn Hilbert*, Math. Ann., 68 (1910), pp. 361–366.

SIAM J. Discrete Math.
Vol. 17, No. 2, pp. 219–236

EXTREMAL SETS MINIMIZING DIMENSION-NORMALIZED BOUNDARY IN HAMMING GRAPHS*

M. CEMIL AZIZOĞLU[†] AND ÖMER EĞECIOĞLU[†]

Abstract. We prove that the set of first k vertices of a Hamming graph in reverse-lexicographic order constitutes an extremal set minimizing the dimension-normalized edge-boundary over all k-vertex subsets of the graph. This generalizes a result of Lindsey and can be used to prove a tight lower bound for the isoperimetric number and the bisection width of arrays.

Key words. Hamming graph, product graph, array, isoperimetric number, bisection width, extremal set, partition, Schur-convexity, majorization

AMS subject classifications. 05C35, 05C40, 05D05

DOI. 10.1137/S0895480100375053

1. Introduction.

We consider questions of the following general form: Given a graph G and a natural number k, what is the optimum value of a certain quantity in a set of k vertices of G? The desired quantity could be the number of edges between a set of k vertices and its complement (i.e., the size of the boundary) or the number of edges induced by a set of k vertices, etc. The sets achieving the optimum value are called *extremal sets*.

Specifically, we study extremal sets in Hamming graphs minimizing the size of the edge-boundary of a set of vertices of given size, where boundary edges along each dimension are *normalized* by a weight determined by that dimension, as shall soon be explained.

First, we introduce some notation and terminology. Given a graph G and a subset X of its vertices, let ∂X denote the *edge-boundary*, or simply *boundary*, of X. This is the set of edges connecting vertices in X with vertices not in X (i.e., the complement of X). A *d-dimensional Hamming graph* H^d is a graph with $k_1 \times k_2 \times \cdots \times k_d$ vertices, $k_1 \leq k_2 \leq \cdots \leq k_d$, each having a unique label $l = \langle l_1, l_2, \ldots, l_d \rangle$, where $0 \leq l_i \leq k_i - 1$. There is an edge between two vertices iff their labels differ in exactly one digit. A *d-dimensional array* A^d resembles H^d with the exception that two vertices are adjacent iff their labels differ in exactly one digit *and* the difference is exactly one. Examples of a two-dimensional Hamming graph and a two-dimensional array are shown in Figure 1.

The *Cartesian product* $G \times H$ of two graphs G and H is the graph with vertex set $V(G) \times V(H)$, in which vertices (u, v) and (u', v') are adjacent iff u is adjacent to u' in G and $v = v'$, or v is adjacent to v' in H and $u = u'$. The constituent graphs G and H are called *factors*. A Hamming graph can be characterized as the Cartesian product of a number of *complete graphs* of different sizes, i.e., $H^d = K_{k_1} \times K_{k_2} \times \cdots \times K_{k_d}$, where K_r is a complete graph on r vertices. Similarly, A^d can be characterized as the Cartesian product of a number of *path graphs* of varying length, i.e., $A^d = P_{k_1} \times P_{k_2} \times \cdots \times P_{k_d}$, where P_r is a path graph (chain) with r vertices.

*Received by the editors July 10, 2000; accepted for publication (in revised form) May 30, 2003; published electronically November 4, 2003.

http://www.siam.org/journals/sidma/17-2/37505.html

[†]Department of Computer Science, University of California at Santa Barbara, Santa Barbara, CA 93106 (azizoglu@cs.ucsb.edu, omer@cs.ucsb.edu). The first author was supported in part by NSF grant CCR–9821038 and by a fellowship from The İzmir Institute of Technology, İzmir, Turkey. The second author was supported in part by NSF grant CCR–9821038.

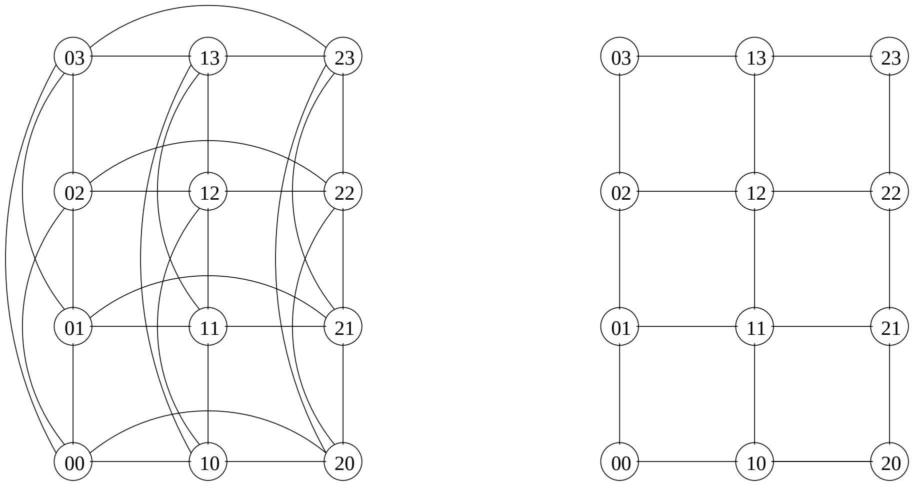


FIG. 1. *The two-dimensional Hamming graph $K_3 \times K_4$ and array $P_3 \times P_4$.*

Lindsey [19] proved that the set of first k vertices of a Hamming graph in *lexicographic order* constitutes an extremal set minimizing the boundary ∂X over all k-element subsets X. The lexicographic order is defined as follows: In the Hamming graph $H^d = K_{k_1} \times K_{k_2} \times \cdots \times K_{k_d}$ with $k_1 \leq k_2 \leq \cdots \leq k_d$, vertex $x = \langle x_1, \ldots, x_d \rangle$ precedes vertex $y = \langle y_1, \ldots, y_d \rangle$ in lexicographic order iff there exists an index i such that $x_1 = y_1$, $x_2 = y_2$, $\ldots$, $x_{i-1} = y_{i-1}$ and $x_i < y_i$ holds. Intuitively, in lexicographic order, we traverse the Hamming graph in the direction of the *next largest factor* starting with the vertex labeled $\langle 0, 0, \ldots, 0 \rangle$. For instance, the vertices of the Hamming graph in Figure 1 in lexicographic order are labeled

$$00,\ 01,\ 02,\ 03,\ 10,\ 11,\ 12,\ 13,\ 20,\ 21,\ 22,\ 23.$$

Our aim in this paper is to determine and describe extremal sets of Hamming graphs minimizing the dimension-normalized boundary. This is defined next.

DEFINITION 1.1. *Given a Hamming graph $H^d = K_{k_1} \times K_{k_2} \times \cdots \times K_{k_d}$ and a subset X of its vertices,* the dimension-normalized boundary $B(X)$ *of X is defined as*

$$B(X) = \frac{|\partial_1 X|}{c_1} + \frac{|\partial_2 X|}{c_2} + \cdots + \frac{|\partial_d X|}{c_d}, \tag{1.1}$$

where for $1 \leq i \leq d$, $\partial_i X$ is the set of boundary edges along dimension i and

$$c_i = \begin{cases} k_i^2 & \text{if } k_i \text{ is even,} \\ k_i^2 - 1 & \text{if } k_i \text{ is odd.} \end{cases} \tag{1.2}$$

We prove that the set of first k vertices in *reverse-lexicographic order* constitutes an extremal set minimizing the dimension-normalized boundary over all k-element subsets in a Hamming graph. The definition of the reverse-lexicographic order is similar to that of the lexicographic order: In the Hamming graph $H^d = K_{k_1} \times K_{k_2} \times \cdots \times K_{k_d}$ with $k_1 \leq k_2 \leq \cdots \leq k_d$, vertex $x = \langle x_1, \ldots, x_d \rangle$ precedes vertex $y = \langle y_1, \ldots, y_d \rangle$ in reverse-lexicographic order iff there exists an index i such that $x_d = y_d$, $x_{d-1} = y_{d-1}$, $\ldots$, $x_{i+1} = y_{i+1}$ and $x_i < y_i$ holds. In other words, we move in the direction of the *next smallest factor* starting at the vertex labeled $\langle 0, 0, \ldots, 0 \rangle$. To illustrate, the vertices of the Hamming graph in the above example listed in reverse-lexicographic order are

$$00,\ 10,\ 20,\ 01,\ 11,\ 21,\ 02,\ 12,\ 22,\ 03,\ 13,\ 23.$$

We should point out that there are other sets of vertices which are structurally equivalent to the sets specified in our definitions of lexicographic or reverse-lexicographic orders. These are obtained by symmetries in the underlying graph. For instance, in Figure 1, another ordering structurally equivalent to the lexicographic ordering would be 23, 22, 21, 20, 13, etc. Similarly, the sets defined by the initial segments of the ordering 23, 13, 03, 22, 12, etc., give rise to sets structurally identical to those in reverse-lexicographic order.

We state our claim formally in the following theorem.

THEOREM 1.2. *Given a d-dimensional Hamming graph H^d, let X be any k-vertex subset of $V(H^d)$ and $\overline{X}$ be the set of first k vertices of H^d in reverse-lexicographic order. Then $B(\overline{X}) \leq B(X)$.*

Interestingly, when all factors of H^d have equal size, the lexicographic and reverse-lexicographic orders both result in structurally symmetric subsets and hence are equivalent with respect to extremal sets minimizing the boundary (dimension-normalized or otherwise). Therefore Theorem 1.2 is trivially true when $k_1 = k_2 = \cdots = k_d$ by Lindsey's result, since the denominators c_i in (1.1) will all be equal and minimizing $B(X)$ will be equivalent to minimizing $|\partial_1 X| + |\partial_2 X| + \cdots + |\partial_d X| = |\partial X|$, i.e., the size of the boundary of X.

In the next section, we describe the notion of the isoperimetric number, which is a quantity closely related to extremal sets. The isoperimetric number problem for special classes of graphs provides the basis of our motivation for this work.

1.1. Motivation. An important quantity in the theory of graphs is the *isoperimetric number* $i(G)$ of a graph G, defined as

$$i(G) = \min_{1 \leq |X| \leq \frac{|V(G)|}{2}} \frac{|\partial X|}{|X|}, \tag{1.3}$$

where $X \subseteq V(G)$. That is, the set of vertices of G is partitioned into two nonempty sets and the ratio of the number of edges between the two parts and the number of vertices in the smaller one is minimized. A subset X achieving the equality in (1.3) is called an *isoperimetric set*.

The notion of the isoperimetric number of a graph G serves as a measure of connectivity of G as it quantifies the minimal interaction between a set of vertices X and its complement $V(G) \setminus X$ in terms of the number of edges between them. In many instances, the isoperimetric number of a graph can be used to obtain a tight lower bound for its *bisection width* as well [18]. We refer the reader to Mohar [22] or Chung [12] for a discussion of basic results and various interesting properties of $i(G)$.

At present, the isoperimetric number of an array $A^d = P_{k_1} \times P_{k_2} \times \cdots \times P_{k_d}$ is known only when either $k_1 = k_2 = \cdots = k_d$ (see Azizoğlu and Eğecioğlu [4]) or the size of the largest factor is even (see Azizoğlu and Eğecioğlu [5]). The latter is also implicit in [10] (see also [17]). See also [3] and [13]. The techniques used to obtain these results seem to fail in the general case. However, using the notion of extremal sets minimizing dimension-normalized boundary together with a result of Nakano [23], one can show that

$$i(A^d) = \min_i \frac{1}{\lfloor \frac{k_i}{2} \rfloor}.$$

The technique used involves embedding a Hamming graph into A^d and associating these extremal sets with isoperimetric sets of the array. We refer the reader to [6] for details.

1.2. A summary of previous results. There has been a significant amount of research in the area of isoperimetric bounds on various popular classes of graphs such as Hamming graphs, arrays, and tori. We shall only mention those results in this area which pertain to our discussion and refer the reader to Bezrukov [8] for a comprehensive survey and Bollobás [9] for a general discussion of this and related topics.

As mentioned before, an extremal set of a graph for a given k is, in a broad sense, a configuration of k vertices with

- minimum number of boundary edges or
- maximum number of spanned edges

among all such k-vertex subsets of the given graph. The problem of finding extremal sets of the first (or second) type is called *the minimum-boundary-edge problem* (or *the maximum-induced-edge problem*). It can be shown that the minimum-boundary-edge and the maximum-induced-edge problems are equivalent for regular graphs [11]. We remark that one can easily obtain the isoperimetric number of a given graph if the extremal sets of the first type are known (and the boundary is actually computable). Evidently, an extremal set X with $\lfloor |V(G)|/2 \rfloor$ vertices in a given graph G determines a bisection for G.

The maximum-induced-edge problem (hence the minimum-boundary-edge problem, because of its regularity) for the hypercube (d-dimensional binary Hamming graph) was solved by Harper [14] and extended by Lindsey [19] to the d-dimensional k-ary Hamming graph. In both instances, there is a nested structure of solutions, and the first k vertices in *lexicographic order* constitute an extremal set. The maximum-induced-edge problem for the d-dimensional k-ary array A_k^d was solved by Bollobás and Leader [11]. Since A_k^d is not regular, this result does not automatically give a solution to the minimum-boundary-edge problem. It was later extended to general arrays by Ahlswede and Bezrukov [1] who also gave a solution for $P_{k_1} \times P_{k_2}$ for the minimum-boundary-edge problem. The first nontrivial bounds on the minimum-boundary-edge problem for the d-dimensional k-ary arrays are in Bollobás and Leader [11]. Unfortunately, however, the bounds obtained are not tight enough to yield an exact formula for $i(A_k^d)$.

Similar problems have been studied in the literature for the vertex-boundary of a given configuration of vertices. For instance, for the d-dimensional k-ary torus, Bollobás and Leader [10] solved the vertex-boundary problem for even k. Riordan [24] later extended their result by giving an ordering of vertices on the d-dimensional even torus, which minimizes the number of vertices at shortest distance t from the vertices in the ordering. Wang and Wang [25] solved this problem for $P_\infty \times \cdots \times P_\infty$, i.e., the *d-dimensional infinite array*, where the minimum is taken over all nonempty finite subsets of vertices. In their result, each P_∞ may be infinite in both directions or in one direction only. They also gave a simple ordering of the vertices in which the first k vertices constitute an extremal set minimizing the vertex-boundary. In a recent paper, Harper [15] solved the vertex-boundary problem on Hamming graphs.

1.3. Outline. The outline of the remainder of this paper is as follows. In section 2 we consider the case of two-dimensional Hamming graphs. First we define the terminology we use and state a number of basic facts on restricted integer partitions, majorization, and Schur-convexity. Then we identify potential extremal sets in H^2 as integer partitions inside a rectangle. The problem of showing that the set of first k vertices of H^2 in reverse-lexicographic order constitutes an extremal set minimizing the dimension-normalized edge-boundary over all k-vertex subsets becomes the problem of maximization of a certain function on partitions, which is a linear combination of two Schur-convex functions. However, the function itself is not Schur-convex, and the identification of the partition on which the maximum is achieved is actually done using an inductive argument. The main result of this section is Lemma 2.5. In section 3 we extend the proof to the higher-dimensional case. This is done by an induction on the number of dimensions, using the two-dimensional result as the base case. Finally, concluding remarks are given in section 4.

2. The two-dimensional case. Let $H^2 = K_m \times K_n$ be a given two-dimensional Hamming graph. Without loss of generality, we may assume that $m = k_1 \leq k_2 = n$. Consider a subset X of vertices in H^2. Let X' be the subset of vertices of H^2 obtained by pushing (compressing) all the vertices in X as far downward and then to the left in H^2 as possible. It is easy to see (and proved in [19], [16]) that $B(X') \leq B(X)$ since the number of boundary edges in either dimension will not increase as a result of this procedure. A subset X' in the compressed form corresponds to a *partition* of the integer $|X|$ contained in the $m \times n$ rectangle.

We give below the definitions and properties of partitions that we will use in our proof of Theorem 1.2. The reader is referred to [2] for further details.

Partitions. A *partition* λ of an integer N is a sequence $(\lambda_1, \lambda_2, \ldots, \lambda_\ell)$ of positive integers (called *parts*) satisfying $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_\ell$ and $\lambda_1 + \lambda_2 + \cdots + \lambda_\ell = N$. We put $|\lambda| = N$. The Ferrers diagram of λ is a two-dimensional array of unit cells (or nodes) in which row i from the bottom has λ_i cells and the rows are left justified. It is clear that, in our case, $\lambda = X'$ forms a partition of $|X|$ whose Ferrers diagram is contained in the $m \times n$ rectangle, i.e., $\lambda_i \leq m$ for $1 \leq i \leq \ell$ (i.e., each part at most m) and $\ell \leq n$ (i.e., number of parts at most n). We use $\mathbb{P}(m,n)$ to denote the set of these partitions. Thus we may assume that an extremal set is a partition $\lambda \in \mathbb{P}(m,n)$, and we use the symbol $\mathbb{P}(m,n)$ to refer to $H^2 = K_m \times K_n$ when we are not interested in the graph structure of H^2 but just the placement of the subset λ. We may augment partitions by adding parts of zero length and write $\sum_{i\geq 1} \lambda_i$ for $|\lambda|$. We also identify partitions with their diagrams when there is no confusion.

Given a partition $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$, we may define a new partition $\lambda' = (\lambda'_1, \lambda'_2, \ldots, \lambda'_m)$ by choosing λ'_i as the number of parts of λ that are $\geq i$. The partition λ' is called the *conjugate* of λ. Geometrically, λ' is obtained from λ by reflection in the main diagonal (equivalently by counting the cells in successive columns of λ). For example, the conjugate of $(5,4,3,3,1,1)$ is $(6,4,4,2,1)$. Clearly $|\lambda| = |\lambda'|$, and if $\lambda \in \mathbb{P}(m,n)$, then $\lambda' \in \mathbb{P}(n,m)$.

Durfee square. Let $d = d(\lambda)$ denote the number of λ_i such that $\lambda_i \geq i$. Then d measures the largest square of cells contained in the partition λ, i.e., the number of cells on the main diagonal of λ, the cells with coordinates of the form (i,i). This square is called the *Durfee square*, and d is called the side of the Durfee square. For the partition $\lambda = (5,4,3,3,1,1)$, the side of the Durfee square is $d = 3$.

Frobenius notation. Suppose d is the side of the Durfee square of λ. Let $\alpha_i = \lambda_i - i$ be the number of cells in the ith row of λ to the right of (i,i) for $1 \leq i \leq d$, and let $\beta_i = \lambda'_i - i$ be the number of cells in the ith column of λ above (i,i) for $1 \leq i \leq d$. Then we have $\alpha_1 > \alpha_2 > \cdots > \alpha_d \geq 0$ and $\beta_1 > \beta_2 > \cdots > \beta_d \geq 0$. The

Frobenius notation for λ is

$$\lambda = (\alpha_1, \ldots, \alpha_d | \beta_1, \ldots, \beta_d) = (\alpha|\beta).$$

For example, if $\lambda = (5,4,3,3,1,1)$, then $\alpha = (4,2,0)$ and $\beta = (5,2,1)$ as shown in Figure 2.

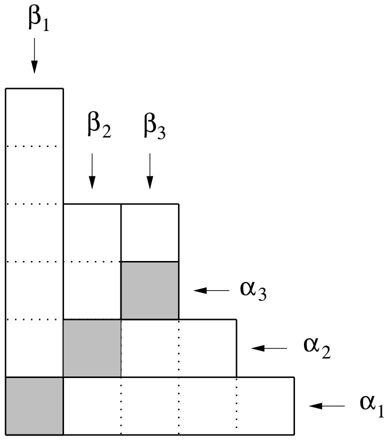


FIG. 2. *The main diagonal (cells in dark) of the* 3×3 *Durfee square and* $\alpha = (4,2,0)$, $\beta = (5,2,1)$ *of the Frobenius notation for the partition* $\lambda = (5,4,3,3,1,1)$.

Reverse-lexicographic ordering on partitions. Given partitions λ and μ, μ *precedes* λ in reverse-lexicographic ordering, denoted by $\mu \geq \lambda$, if either $\lambda = \mu$ or else the first nonvanishing difference $\lambda_i - \mu_i$ is positive. Reverse-lexicographic ordering is a total order. For example, partitions of $N = 5$ are ordered by reverse-lexicographic ordering as

$$(5) \geq\ (4,1) \geq\ (3,2) \geq\ (3,1,1) \geq\ (2,2,1) \geq\ (2,1,1,1) \geq\ (1,1,1,1,1),$$

the first (or the "smallest" one) being (5). The reason for this reversed notation is for consistency with the *dominance order* on partitions that we later define.

Majorization, Schur-convexity, and transfer. Given two partitions $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_N)$ of N, λ is *majorized* by μ, written $\lambda \prec \mu$, if

$$\lambda_1 + \lambda_2 + \cdots + \lambda_k \leq \mu_1 + \mu_2 + \cdots + \mu_k, \qquad k = 1, 2, \ldots, N.$$

Majorization is also referred to as the *dominance* or *natural* order [20, Chap. 1]. As soon as $N \geq 6$, majorization is not a total ordering. For example, the partitions $(3,1,1,1)$ and $(2,2,2)$ of 6 are not comparable. However, reverse-lexicographic ordering on partitions is a linear extension of $\prec$. Thus

$$\lambda \prec \mu \ \Rightarrow\ \lambda \leq \mu.$$

Furthermore $\lambda \prec \mu \Leftrightarrow \mu' \prec \lambda'$ (see [20, (1.11)]). A real-valued function g defined on partitions of an integer N is said to be *Schur-convex* (see [21, Chap. 3]) if

$$\lambda \prec \mu \ \Rightarrow\ g(\lambda) \leq g(\mu).$$

We make use of the following special case of a result of Schur, 1923, and Hardy, Littlewood, and Polya, 1929 (see [21, Chap. 3, Prop. C.1]).

PROPOSITION 2.1. *Suppose ϕ is a real-valued convex function on $\mathbb{R}$ and N is a positive integer. Then the function*

$$g(\lambda) = \sum_{i \geq 1} \phi(\lambda_i)$$

is Schur-convex on partitions of N.

Given a partition $\mu = (\mu_1, \mu_2, \ldots, \mu_N)$ with $\mu_i > \mu_j$, the transformation that takes μ to $\rho = (\rho_1, \rho_2, \ldots, \rho_N)$ defined by

$$\begin{aligned} \rho_i &= \mu_i - 1, \\ \rho_j &= \mu_j + 1, \\ \rho_k &= \mu_k, \qquad\qquad k \neq i, j, \end{aligned}$$

is called a *transfer* from i to j. By a result of Muirhead, if $\lambda \prec \mu$, then λ can be derived from μ by successive application of a finite number of transfers [21, Chap. 5, D.1], [20, (1.16)].

Now consider a partition $\lambda \in \mathbb{P}(m, n)$, where $k_1 = m \leq n = k_2$, which corresponds to a compressed set in $H^2 = K_m \times K_n$. Let $\partial_m \lambda$ and $\partial_n \lambda$ be sets of horizontal and vertical boundary edges of λ, respectively. Then we have

$$|\partial_m \lambda| = \sum_{\lambda_i > 0} \lambda_i (m - \lambda_i) \qquad \text{and} \qquad |\partial_n \lambda| = \sum_{\lambda'_j > 0} \lambda'_j (n - \lambda'_j).$$

After substituting these into (1.1) and eliminating constant terms, we see that finding a subset $\lambda \in K_m \times K_n$ minimizing $B(\lambda)$ is equivalent to *maximizing* the following function f:

$$\begin{aligned} f(\lambda) &= c_1 \sum_{i=1}^{\lambda'_1} \lambda_i^2 + c_2 \sum_{j=1}^{\lambda_1} \lambda_j'^2 \\ &= \gamma_n \sum_{i \geq 1} \lambda_i^2 + \gamma_m \sum_{j \geq 1} \lambda_j'^2 \end{aligned} \tag{2.1}$$

on $\mathbb{P}(m, n)$ $(m \leq n)$, where

$$\gamma_n = \begin{cases} n^2 & \text{if } n \text{ is even,} \\ n^2 - 1 & \text{if } n \text{ is odd} \end{cases} \tag{2.2}$$

in accordance with the definition of the weights c_i in (1.2). We prove the following equivalent formulation of Theorem 1.2 for $H^2 = K_m \times K_n$:

THEOREM 2.2. *When restricted to partitions of a fixed $N \leq mn$, the function f defined in* (2.1) *is maximized on $\mathbb{P}(m, n)$, $m \leq n$, by the reverse-lexicographically smallest partition of N in $\mathbb{P}(m, n)$.*

The proof of the main result of this paper, and consequently the proof of the formula for the isoperimetric number of arrays itself (see [6]) which uses this result, would be simplified by an independent proof of this fact. However, the function f is not Schur-convex. In other words, transfer operators [21, Chap. 5, D.1] or equivalently raising/lowering operators (see [20, (1.15)–(1.16)]) which move from a given λ to a smaller one in the linear order while keeping the value of f nondecreasing are insufficient to prove this fact. As an example take $m = 4$, $n = 8$, $\lambda = (4, 1, 1, 1, 1, 1, 1)$, $\mu = (4, 2, 1, 1, 1, 1)$. Then $\lambda \prec \mu$ by a single transfer as shown in Figure 3, but $2240 = f(\lambda) > f(\mu) = 2208$. It can be shown that transfer arguments can be used to prove Theorem 2.2 in the special case when $n \geq m^2$.

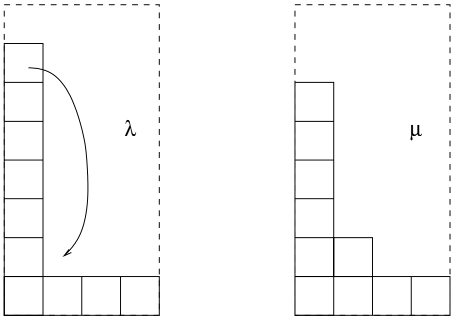


FIG. 3. *μ is obtained from λ by single transfer (lowering the indicated cell). Here $m = 4$, $n = 8$, $\lambda \prec \mu$ but $f(\lambda) = 2240$, whereas $f(\mu) = 2208$.*

We reformulate $f(\lambda)$ in (2.1) in a form which is more convenient for our characterization. Given a partition $\lambda \in \mathbb{P}(m, n)$, let $\widetilde{\lambda}$ be the partition in $\mathbb{P}(m - 1, n - 1)$ which is obtained by removing the bottommost row and the leftmost column from the $m \times n$ rectangle. Now consider the term $\gamma_n \sum \lambda_i^2$ of $f(\lambda)$ in (2.1). By taking the first term λ_1^2 out of the summation and putting $\lambda_i = (\lambda_i - 1) + 1$, we have

$$\begin{aligned}\gamma_n \sum_{i\geq 1} \lambda_i^2 &= \gamma_n \left[\lambda_1^2 + \sum_{i=2}^{\lambda_1'} ((\lambda_i - 1) + 1)^2\right] \\ &= \gamma_n \left[\lambda_1^2 + \sum_{i\geq 1} \widetilde{\lambda}_i^2 + 2\sum_{i\geq 2} \lambda_i - \sum_{i=2}^{\lambda_1'} 1\right] \\ &= \gamma_n \left[\lambda_1^2 + \sum_{i\geq 1} \widetilde{\lambda}_i^2 + 2(|\lambda| - \lambda_1) - (\lambda_1' - 1)\right].\end{aligned}$$

Putting the second term $\gamma_m \sum_{j\geq 1} \lambda_j'^2$ of $f(\lambda)$ as above, we now have

$$\begin{aligned}f(\lambda) = \gamma_n &\left[\lambda_1^2 + \sum_{i\geq 1} \widetilde{\lambda}_i^2 + 2(|\lambda| - \lambda_1) - (\lambda_1' - 1)\right] \\ &+ \gamma_m \left[\lambda_1'^2 + \sum_{j\geq 1} \widetilde{\lambda}_j'^2 + 2(|\lambda| - \lambda_1') - (\lambda_1 - 1)\right].\end{aligned}$$

Since $f(\widetilde{\lambda}) = \gamma_n \sum_{i\geq 1} \widetilde{\lambda}_i^2 + \gamma_m \sum_{j\geq 1} \widetilde{\lambda}_j'^2$, this is equivalent to

$$\begin{aligned}f(\lambda) = f(\widetilde{\lambda}) &+ \gamma_n \left[\lambda_1^2 + 2(|\lambda| - \lambda_1) - (\lambda_1' - 1)\right] \\ &+ \gamma_m \left[\lambda_1'^2 + 2(|\lambda| - \lambda_1') - (\lambda_1 - 1)\right].\end{aligned} \tag{2.3}$$

Suppose now $\lambda = (\alpha|\beta) = (\alpha_1, \ldots, \alpha_d|\beta_1, \ldots, \beta_d)$. Then the partition $\widetilde{\lambda}$ obtained from λ by deleting the leftmost column and the bottommost row of the $m \times n$ rectangle is a partition $(\widetilde{\alpha}|\widetilde{\beta})$ which has a Durfee square of size $d-1$, where $\widetilde{\lambda} = (\widetilde{\alpha}|\widetilde{\beta}) = (\alpha_2, \ldots, \alpha_d|\beta_2, \ldots, \beta_d)$. Using this and $\lambda_1 = 1 + \alpha_1$, $\lambda'_1 = 1 + \beta_1$, equality (2.3) can be reformulated as

$$f(\alpha|\beta) = f(\widetilde{\alpha}|\widetilde{\beta}) + \gamma_n \alpha_1^2 - \gamma_m \alpha_1 + \gamma_m \beta_1^2 - \gamma_n \beta_1 + (\gamma_n + \gamma_m)(2|\lambda| - 1). \tag{2.4}$$

Since the last term $(\gamma_n + \gamma_m)(2|\lambda| - 1)$ is constant for all configurations with size $|\lambda|$, iterating this expression we have the following proposition.

PROPOSITION 2.3. *Over partitions $\lambda = (\alpha|\beta)$ in $\mathbb{P}(m,n)$ of a fixed integer $|\lambda|$ with Durfee square of size d, maximizing $f(\lambda)$ is equivalent to maximizing*

$$\gamma_n \left[\alpha_1^2 + \cdots + \alpha_d^2 - (\beta_1 + \cdots + \beta_d)\right] + \gamma_m \left[\beta_1^2 + \cdots + \beta_d^2 - (\alpha_1 + \cdots + \alpha_d)\right]. \tag{2.5}$$

Durfee-equivalence. Suppose $\lambda \in \mathbb{P}(m,n)$ has a Durfee square D of size d. Let $\nu = \nu(\lambda)$ denote the partition that lies north (on top) of D and $\eta = \eta(\lambda)$ the partition that lies east (to the right) of D. Then $\nu_1 \le d$ and $\nu'_1 \le n - d$, and $\eta'_1 \le d$ and $\eta_1 \le m - d$. Two partitions $\lambda, \mu \in \mathbb{P}(m,n)$ are *Durfee-equivalent* iff

1. $d(\lambda) = d(\mu)$,
2. $|\nu(\lambda)| = |\nu(\mu)|$ and $|\eta(\lambda)| = |\eta(\mu)|$.

We single out a special representative λ^* in the equivalence class of partitions Durfee-equivalent to λ. λ^* is the partition in which η is the largest in the dominance order in the $d \times (m-d)$ rectangle to the right of the Durfee square and ν' is the largest in the dominance order in the $(n-d) \times d$ rectangle to the top of the Durfee square. In other words, in λ^*, η^* is obtained by distributing $|\eta|$ cells into as many rows as possible of length $m-d$, followed by a (possibly null) partial row of size r. Similarly in λ^*, ν^* is obtained by distributing $|\nu|$ cells by first laying as many columns as possible of length $n-d$, followed by a (possibly null) partial column of size s. An example of this is shown in Figure 4.

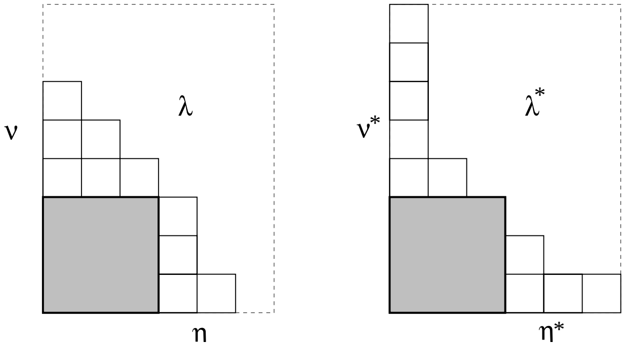


FIG. 4. *Partition $\lambda = (5,4,4,3,2,1)$ is Durfee-equivalent to the special representative $\lambda^* = (6,4,3,2,1,1,1,1)$ in $\mathbb{P}(6,8)$.*

PROPOSITION 2.4. *Suppose $\lambda \in \mathbb{P}(m,n)$ and λ^* is the special representative of λ in the equivalence class of partitions Durfee-equivalent to λ. Then $f(\lambda^*) \ge f(\lambda)$.*

Proof. We use Proposition 2.3. Since in Durfee-equivalence $|\nu|$ and $|\eta|$ do not change, $\alpha_1 + \cdots + \alpha_d$ and $\beta_1 + \cdots + \beta_d$ are constant. Thus maximizing f over the

Durfee-equivalence class of λ is equivalent to maximizing

$$\gamma_n \left(\alpha_1^2 + \cdots + \alpha_d^2\right) + \gamma_m \left(\beta_1^2 + \cdots + \beta_d^2\right),$$

which is decoupled. Since the function $\phi(x) = x^2$ is convex on $\mathbb{R}$, applying the majorization result of Proposition 2.1 to each term separately, we obtain the proposition. □

Remark. Proposition 2.4 allows us to restrict potential maximizers of the function $f(\lambda)$ on $\lambda \in \mathbb{P}(m,n)$ to partitions of the form shown in Figure 5. Here $|\lambda| = d^2 + w(m-d) + t(n-d) + r + s$.

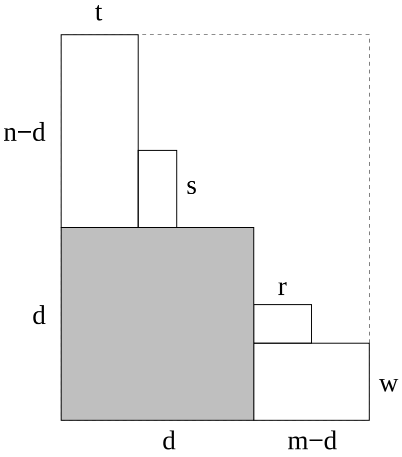


FIG. 5. *The form of special representatives of Durfee-equivalence classes of partitions in* $\mathbb{P}(m,n)$.

2.1. Extremal sets for the two-dimensional Hamming graph. Now we are ready to prove the two-dimensional case, which is stated using the terminology of Hamming graphs in the following lemma.

LEMMA 2.5. *Given a two-dimensional Hamming graph $H^2 = K_m \times K_n$ with $m \le n$, let λ be any k-vertex subset of $V(H^2)$ and $\overline{\lambda}$ be the set of first k vertices of H^2 in reverse-lexicographic order. Then $f(\overline{\lambda}) \ge f(\lambda)$. That is,*

$$\gamma_n \sum_{i\ge1} \overline{\lambda}_i^2 + \gamma_m \sum_{j\ge1} \overline{\lambda}_j'^2 \ge \gamma_n \sum_{i\ge1} \lambda_i^2 + \gamma_m \sum_{j\ge1} \lambda_j'^2. \tag{2.6}$$

Proof. We give the proof only for n and m both even. The other cases are similar. By Proposition 2.4, we can assume that $\lambda = \lambda^*$ is the special representative in the Durfee-equivalence class of λ and is characterized by the parameters r, s, w, t, m, d, n as shown in Figure 5 with $|\lambda| = d^2 + w(m-d) + t(n-d) + r + s$. Using the original definition (2.1) of f, we compute

$$\gamma_n \sum \lambda_i^2 = n^2 \Big[wm^2 + (d+r)^2 + (d-w-1)d^2 + s(t+1)^2 + (n-d-s)t^2\Big],$$

$$\gamma_m \sum_{j\ge1} \lambda_j'^2 = m^2 \Big[tn^2 + (d+s)^2 + (d-t-1)d^2 + r(w+1)^2 + (m-d-r)w^2\Big].$$

For simplicity, assume that m divides $|\lambda|$. Then $\overline{\lambda}$ consists of $|\lambda|/m$ rows of length m each. Thus

$$\begin{aligned}\gamma_n \sum \overline{\lambda}_i^2 &= n^2 m \left[d^2 + w(m-d) + t(n-d) + r + s\right],\\ \gamma_m \sum_{j\geq 1} \overline{\lambda}_j'^2 &= m \left[d^2 + w(m-d) + t(n-d) + r + s\right]^2.\end{aligned}$$

Let $g(r,s,w,t,m,d,n) = f(\overline{\lambda}) - f(\lambda)$. Then

$$\begin{aligned}(2.7)\quad & g(r,s,w,t,m,d,n)\\ &= n^2 m \left[d^2 + w(m-d) + t(n-d) + r + s\right]\\ &\quad + m \left[d^2 + w(m-d) + t(n-d) + r + s\right]^2\\ &\quad - n^2 \left[wm^2 + (d+r)^2 + (d-w-1)d^2 + s(t+1)^2 + (n-d-s)t^2\right]\\ &\quad - m^2 \left[tn^2 + (d+s)^2 + (d-t-1)d^2 + r(w+1)^2 + (m-d-r)w^2\right].\end{aligned}$$

g is a polynomial of total degree 5 in the integer variables r, s, w, t, m, d, n, which is quadratic as a polynomial in r, s, w, t, and m, cubic in n, and quartic in d. Let R be region defined by the inequalities

$$(2.8)\qquad \begin{aligned}&0 \leq r \leq m-d,\\ &0 \leq s \leq n-d,\\ &0 \leq w \leq d-1,\\ &0 \leq t \leq d-1,\\ &d \leq m \leq n\end{aligned}$$

that we read off from Figure 5. Now we show that $g(r,s,w,t,m,d,n) \geq 0$ on R where g is as in (2.7) and R is the region defined in (2.8). Rewrite the inequalities in R in the form

$$\begin{aligned}&r_0 \leq r \leq r_1,\\ &s_0 \leq s \leq s_1,\\ &w_0 \leq w \leq w_1,\\ &t_0 \leq t \leq t_1,\\ &m_0 \leq m \leq m_1\end{aligned}$$

with $r_0 = 0$, $r_1 = m-d$, and $s_0 = 0$, $s_1 = n-d$, etc., up to $m_0 = d$, $m_1 = n$. The idea of the proof is simple in theory: As a quadratic in r, we calculate that the leading coefficient is $m - n^2 \leq 0$. If, in addition, we can show that $g(r_0,s,w,t,m,d,n) \geq 0$ and $g(r_1,s,w,t,m,d,n) \geq 0$ on R, then we would be done. But this requires that we solve two subproblems: We need to show $g(r_0,s,w,t,m,d,n) \geq 0$ and $g(r_1,s,w,t,m,d,n) \geq 0$. Both of these are quadratic in s. If we can show that the leading coefficient in each is ≤ 0 on R *and* if each one evaluated in $s = s_0$ and $s = s_1$ is ≥ 0 on R, then we would be done. Iterating this argument, to prove the claim about the nonnegativity of g on R, it suffices to verify the following two assertions:

1. $g(r_{i_1},s,w,t,m,d,n)$ has leading coefficient ≤ 0 on R as a polynomial in s, $g(r_{i_1},s_{i_2},w,t,m,d,n)$ has leading coefficient ≤ 0 on R as a polynomial in w, $g(r_{i_1},s_{i_2},w_{i_3},t,m,d,n)$ has leading coefficient ≤ 0 on R as a polynomial in t, $g(r_{i_1},s_{i_2},w_{i_3},t_{i_4},m,d,n)$ has leading coefficient ≤ 0 on R as a polynomial in m for all 0-1 vectors (i_1,i_2,i_3,i_4),
2. $g(r_{i_1},s_{i_2},w_{i_3},t_{i_4},m_{i_5},d,n)$ is ≥ 0 on R for each 0-1 vector (i_1,i_2,i_3,i_4,i_5).

Table 1

Leading coefficients of the quadratic terms in g. For example, the entry in row 011 *indicates that the quadratic $g(r_0, s_1, w_1, t, m, d, n) = g(0, n-d, d-1, t, m, n, d)$ in t has the expression $-(n-d)(n^2 - mn + dm) \leq 0$ as the coefficient of t^2.*

$i_1i_2i_3i_4$	Coefficient of the leading term
ϵ	$-(n^2 - m)$
0	$-m(m-1)$
1	$-m(m-1)$
00	$-dm(m-d)$
01	$-dm(m-d)$
10	$-dm(m-d)$
11	$-dm(m-d)$
000	$-(n-d)(n^2 - mn + dm)$
001	$-(n-d)(n^2 - mn + dm)$
010	$-(n-d)(n^2 - mn + dm)$
011	$-(n-d)(n^2 - mn + dm)$
100	$-(n-d)(n^2 - mn + dm)$
101	$-(n-d)(n^2 - mn + dm)$
110	$-(n-d)(n^2 - mn + dm)$
111	$-(n-d)(n^2 - mn + dm)$
0000	$-d^3$
0001	$-d^2 - (d-1)n^2$
0010	$-d$
0011	$-(d-1)(n-d+1)^2 - 1$
0100	$-d^2(d-1) - n^2$
0101	$-dn^2$
0110	$-(d-1) - (n-d+1)^2$
0111	$-d(n-d+1)^2$
1000	$-d(d-1)^2$
1001	$-(d-1)(n^2 - 2n + d)$
1010	0
1011	$-(d-1)(n-d)^2$
1100	$-d((d-1)(d-2)+1) - n(n-2)$
1101	$-d(n-1)^2$
1110	$-(n-d)^2$
1111	$-d(n-d)^2$

This is a job best suited to a symbolic algebra package. The expressions proving this proposition are given in Tables 1 and 2. They were calculated by a short Mathematica program. □

3. The higher-dimensional case. In this section, we prove Theorem 1.2 for an arbitrary number of dimensions d. The main idea of the proof is based on that of [16]; hence our notation is similar to the notation therein.

Proof. The proof is by induction on d with $d = 2$, already proved in Lemma 2.5, being the base case. We assume $k_1 \leq k_2 \leq \cdots \leq k_d$ for $H^d = K_{k_1} \times K_{k_2} \times \cdots \times K_{k_d}$ and vertices are labeled by d-tuples $\langle l_1, l_2, \ldots, l_d \rangle$, where $0 \leq l_i \leq k_i - 1$.

The idea is to transform a given arbitrary configuration into one in reverse-lexicographic order so as not to increase the normalized boundary. To aid the readability of the proof, Figure 6 provides a three-dimensional Hamming graph $H^3 = K_4 \times K_5 \times K_{10}$, which illustrates the transformation process.

Given an arbitrary configuration X in H^d, we permute the k_d $(d-1)$-dimensional Hamming subgraphs along dimension d such that successive subgraphs have fewer elements of X. Now we apply the induction hypothesis to each of these subgraphs. Phase (i) in Figure 6 illustrates a configuration obtained after this step. Note that applying this procedure cannot increase $B(X)$ since $|\partial_d X|$ cannot increase and by the

Table 2

The values of the specializations of the quadratic terms in g. For example, the entry in row 01101 *indicates that* $g(r_0, s_1, w_1, t_0, m_1, d, n) = g(0, n-d, d-1, 0, n, n, d) = 2n^2(d-1)(n-d) \geq 0$.

$i_1 i_1 i_3 i_4 i_5$	$g(r_{i_1}, s_{i_2}, w_{i_3}, t_{i_4}, m_{i_5}, d, n)$
00000	0
00001	$d^2 n(n-d)^2$
00010	$(d-1)(n-d)(n^2-dn+d^2)$
00011	$dn(n-1)(n-d)$
00100	0
00101	$dn(n-1)(n-d)$
00110	$(d-1)(n-d)(n^2-dn+d^2)$
00111	$(d-1)^2 n(n-d)^2 + 2n^2(d-1)(n-d)$
01000	$(d-1)(n-d)(n^2-dn+d^2)$
01001	$d(d-1)n(n-d)(n-d+1)$
01010	0
01011	0
01100	$(d-1)(n-d)(n^2-dn+d^2)$
01101	$2n^2(d-1)(n-d)$
01110	0
01111	$d(d-1)n(n-d)(n-d+1)$
10000	0
10001	$d(d-1)n(n-d)(n-d+1)$
10010	$(d-1)(n-d)(n^2-dn+d^2)$
10011	$2(d-1)n^2(n-d)$
10100	0
10101	0
10110	$(d-1)(n-d)(n^2-dn+d^2)$
10111	$d(d-1)n(n-d)(n-d+1)$
11000	$(d-1)(n-d)(n^2-dn+d^2)$
11001	$(d-2)^2 n(n-d)^2 + 2n^2(d-1)(n-d)$
11010	0
11011	$dn(n-1)(n-d)$
11100	$(d-1)(n-d)(n^2-dn+d^2)$
11101	$dn(n-1)(n-d)$
11110	0
11111	$d^2 n(n-d)^2$

induction hypothesis

$$\frac{|\partial_1 X_i|}{c_1} + \cdots + \frac{|\partial_{d-1} X_i|}{c_{d-1}}$$

is smallest for each subgraph i, where X_i is the set of elements of X that are in subgraph i. Not surprisingly, this means that candidate extremal sets in higher dimensions are among *higher-dimensional partitions* (see [2, Chap. 11]), which are contained in the d-dimensional parallelepiped $k_1 \times k_2 \times \cdots \times k_d$. Now we repeat the same steps for subgraphs along dimension $d-1$ as well. This step is illustrated by phase (ii) in Figure 6.

Consider the $(d-2)$-dimensional Hamming subgraphs of H^d when dimensions d and $d-1$ are fixed. We call each such subgraph "complete" iff its vertices are completely contained in X, "incomplete" iff there exists some (but not all) contained in X, and "empty" iff none is in X. We shall show that if there are more than one incomplete subgraph, then these can be combined without increasing $B(X)$. The result of this step is shown by phase (iii) in Figure 6. To this end we give some definitions and develop proper notation.

First, suppose $P^*_{p,q}$ and $P^*_{r,s}$ are sets of vertices of two such incomplete $(d-2)$-

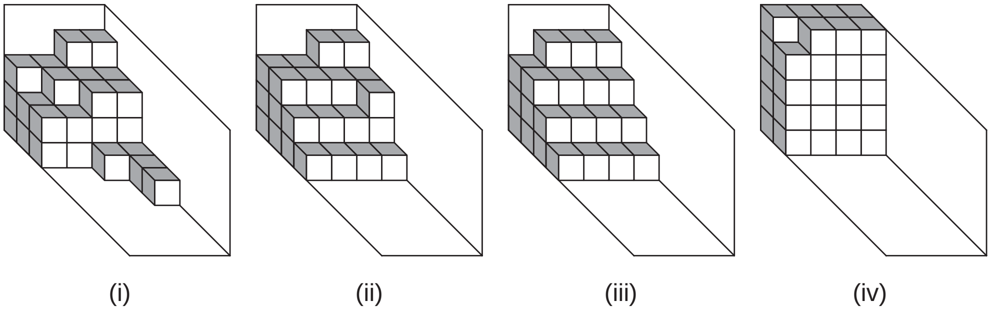


FIG. 6. *Conversion into the reverse-lexicographic order in three dimensions.*

dimensional Hamming subgraphs, where p, r and q, s are coordinates of dimension d and $d-1$, respectively, with $0 \leq p, r \leq k_d - 1$ and $0 \leq q, s \leq k_{d-1} - 1$. Without loss of generality, assume

$$\frac{p}{c_d} + \frac{q}{c_{d-1}} \geq \frac{r}{c_d} + \frac{s}{c_{d-1}}. \tag{3.1}$$

Next, let $P_{p,q} = P^*_{p,q} \cap X$, $P_{r,s} = P^*_{r,s} \cap X$, and $Y = X \setminus (P_{p,q} \cup P_{r,s})$. In our example, there are exactly two such subgraphs with $p = 0$, $q = 3$ and $r = 2$, $s = 2$, i.e., $P_{0,3}$ and $P_{2,2}$, which satisfies the assumption given by the inequality above.

Now given two disjoint subsets S and T of $V(H^d)$, let

$$B(S,T) = \frac{|\partial_1(S,T)|}{c_1} + \frac{|\partial_2(S,T)|}{c_2} + \cdots + \frac{|\partial_d(S,T)|}{c_d}, \tag{3.2}$$

where c_i is as defined before and $\partial_i(S,T)$ is the set of edges in dimension i having one end in S and the other in T. Note that, in this notation, $B(X) = B(X, V(H^d) \setminus X)$.

Note that the following holds:

$$\begin{aligned} B(X) = {} & B(Y) + B(P_{p,q}) + B(P_{r,s}) \\ & - 2(B(Y, P_{p,q}) + B(Y, P_{r,s})) - 2B(P_{p,q}, P_{r,s}). \end{aligned} \tag{3.3}$$

We claim that if as many elements in $P_{r,s}$ as possible are moved to $P^*_{p,q}$ preserving the reverse-lexicographic order, then $B(X)$ does not increase. To this end, consider the terms in (3.3). First, we remark that, by virtue of the reverse-lexicographic order, we must have $p \neq r$ and $q \neq s$, and therefore $B(P_{p,q}, P_{r,s}) = 0$ in (3.3). Furthermore, because of inequality (3.1), $B(Y, P_{p,q}) + B(Y, P_{r,s})$ cannot decrease by this move, and $B(Y)$ is constant. Finally, we claim that $B(P_{p,q}) + B(P_{r,s})$ does not increase.

To prove this, note that any vertex $v \in (P_{p,q} \cup P_{r,s})$ is adjacent to $k_d - 1$ and $k_{d-1} - 1$ vertices in dimensions d and $d-1$, respectively. Thus, moving vertices from $P_{r,s}$ to $P^*_{p,q}$ does not change the boundary along dimensions d and $d-1$. Therefore, it suffices to prove

$$B'(P_{p,q}) + B'(P_{r,s}) \geq B'(P'_{p,q}) + B'(P'_{r,s}), \tag{3.4}$$

where

$$B'(X) = \frac{|\partial_1 X|}{c_1} + \cdots + \frac{|\partial_{d-2} X|}{c_{d-2}}$$

and $P'_{p,q}$ and $P'_{r,s}$ are the new subsets corresponding to $P_{p,q}$ and $P_{r,s}$ respectively, after elements are moved from $P_{r,s}$ to $P^*_{p,q}$.

To prove inequality (3.4), first suppose that all of $P_{r,s}$ fits in the complement of $P_{p,q}$ with respect to $P^*_{p,q}$. Thus we can place elements of $P_{r,s}$ into $P^*_{p,q} \setminus P_{p,q}$ in a set structurally identical to the one given by reverse-lexicographic order, i.e., starting with vertex $\langle k_1-1, k_2-1, \ldots, k_{d-2}-1, q, p\rangle$ of H^d and expanding in the direction of the smallest factor of the Hamming graph. That is,

$$\begin{aligned}\langle k_1-1, k_2-1, \ldots, k_{d-3}-1, k_{d-2}-1, q, p\rangle &\to \langle k_1-1, k_2-1, \ldots, k_{d-3}-1, k_{d-2}-2, q, p\rangle\\ &\to \cdots \to \langle k_1-1, k_2-1, \ldots, k_{d-3}-2, k_{d-2}-1, q, p\rangle\\ &\to \langle k_1-1, k_2-1, \ldots, k_{d-3}-2, k_{d-2}-2, q, p\rangle \to \cdots\end{aligned}$$

and so on. This is shown in Figure 7.

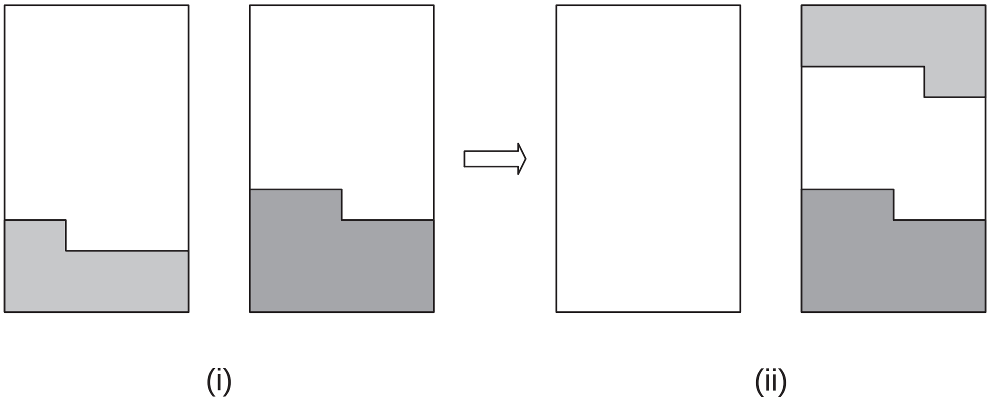


FIG. 7. *Combining two incomplete subgraphs where the elements can fit into one.* (i) *Subgraphs before, and* (ii) *after.*

In this case, we have $B'(P'_{r,s}) = 0$ since $P'_{r,s} = \phi$ and $B'(P'_{p,q})$ can be written as $B'(P'_{p,q}) = B'(P_{p,q}) + B'(P_{r,s}) - 2B'(P_{p,q}, P_{r,s})$. Substituting these values into inequality (3.4), it suffices to prove that

$$B'(P_{p,q}) + B'(P_{r,s}) \geq B'(P_{p,q}) + B'(P_{r,s}) - 2B'(P_{p,q}, P_{r,s}),$$

which obviously holds since $B'(P_{p,q}, P_{r,s}) \geq 0$. We remark that $P'_{p,q}$ is not in reverse-lexicographic order at this point since it consists of two subsets, each of which is structurally in reverse-lexicographic order. Nevertheless, by an easy application of the induction hypothesis, we can convert it to the reverse-lexicographic order without increasing $B'(P'_{p,q})$.

Now assume that not all elements of $P_{r,s}$ fit into $P^*_{p,q}$. First take $|P^*_{p,q} \setminus P_{p,q}|$ vertices in reverse-lexicographic order in $P^*_{r,s}$. These vertices are in $P_{r,s}$. Call this set of vertices Y_2 and set $Y_1 = P_{r,s} \setminus Y_2$. After moving all vertices in Y_2 to $P^*_{p,q}$, we put Y_1 in reverse-lexicographic order $\overline{Y}_1$ within $P^*_{r,s}$. This is shown in Figure 8.

Then, inequality (3.4) reduces to proving

$$B'(P_{p,q}) + B'(P_{r,s}) \geq B'(Y_1)$$

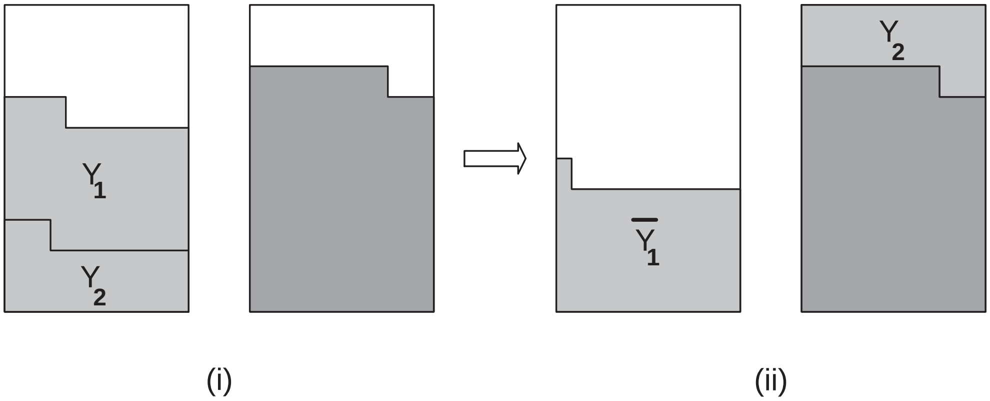


FIG. 8. *Combining two incomplete subgraphs where the elements cannot fit into one.* (i) *Subgraphs before, and* (ii) *after.*

as $B'(Y_1) \geq B'(\overline{Y_1})$ holds by the induction hypothesis. Now note that $B'(P_{r,s}) = B'(Y_1) + B'(Y_2) - 2B'(Y_1, Y_2)$ and $B'(Y_2) = B'(P_{p,q})$ since Y_2 and $P_{p,q}$ are complementary in $P^*_{p,q}$. Thus the above inequality is equivalent to

$$B'(Y_2) \geq B'(Y_1, Y_2),$$

which obviously holds. Thus $B(P_{p,q}) + B(P_{r,s})$ does not increase as claimed. By applying this process to all $(d-2)$-dimensional incomplete subgraphs, we can assume that X has only one incomplete $(d-2)$-dimensional Hamming subgraph.

Finally we treat the $(d-2)$-dimensional Hamming subgraphs as single vertices and use the two-dimensional case to minimize $B(X)$ by putting them in reverse-lexicographic order with the only incomplete one highest in the order, as shown by phase (iv) in Figure 6. This completes the proof of Theorem 1.2. □

4. Conclusions. We proved that the set of first k vertices of the Hamming graph $H^d = K_{k_1} \times K_{k_2} \times \cdots \times K_{k_d}$ $(k_1 \leq k_2 \leq \cdots \leq k_d)$ in reverse-lexicographic order constitutes an extremal set minimizing the dimension-normalized edge-boundary over all k-vertex subsets of the graph. The boundary edges $\partial_i X$ along the ith dimension of $X \subset V(H^d)$ are normalized by a weight

$$c_i = \begin{cases} k_i^2 & \text{if } k_i \text{ is even,} \\ k_i^2 - 1 & \text{if } k_i \text{ is odd,} \end{cases}$$

which naturally arises in the isoperimetric number problem for d-dimensional arrays. The weighted boundary to be minimized is then

$$B(X) = \frac{|\partial_1 X|}{c_1} + \frac{|\partial_2 X|}{c_2} + \cdots + \frac{|\partial_d X|}{c_d}$$

over $X \subset V(H^d)$. Interestingly, when all factors of H^d have equal size, the lexicographic and reverse-lexicographic orders both result in structurally symmetric subsets and hence are equivalent with respect to extremal sets minimizing the boundary (dimension-normalized or otherwise). Thus our result is identical to Lindsey's for $k_1 = k_2 = \cdots = k_d$.

We formulated the problem for the two-dimensional case as the maximization of the function f defined on partitions $\lambda \in \mathbb{P}(m, n)$ $(m \leq n)$ by

$$f(\lambda) = \gamma_n \sum_{i=1}^{n} \lambda_i^2 + \gamma_m \sum_{j=1}^{m} \lambda_j'^2$$

and proved that f is maximized for $N \leq nm$, by the reverse-lexicographically smallest partition of N in $\mathbb{P}(m, n)$, where

$$\gamma_n = \begin{cases} n^2 & \text{if } n \text{ is even,} \\ n^2 - 1 & \text{if } n \text{ is odd.} \end{cases}$$

This result for $d = 2$ forms the base step of the higher-dimensional case.

Acknowledgment. The authors would like to thank the anonymous referee whose careful repeated reviews were essential for us to obtain correct proofs presented in this revised version.

REFERENCES

[1] R. Ahlswede and S. L. Bezrukov, *Edge-isoperimetric theorems for integer point arrays*, Appl. Math. Lett., 8 (1995), pp. 75–80.

[2] G. E. Andrews, *The Theory of Partitions*, Addison–Wesley, Reading, MA, 1976.

[3] M. C. Azizoğlu and Ö. Eğecioğlu, *Isoperimetric number of the Cartesian product of graphs and paths*, Congr. Numer., 131 (1998), pp. 135–143.

[4] M. C. Azizoğlu and Ö. Eğecioğlu, *The isoperimetric number of d–dimensional k–ary arrays*, Internat. J. Found. Comput. Sci., 10 (1999), pp. 289–300.

[5] M. C. Azizoğlu and Ö. Eğecioğlu, *The isoperimetric number and the bisection width of generalized cylinders*, Electronic Notes in Discrete Mathematics, 11 (2002).

[6] M. C. Azizoğlu and Ö. Eğecioğlu, *The bisection width and the isoperimetric number of arrays*, Discrete Appl. Math., in press.

[7] S. L. Bezrukov, *Variational Principle in Discrete Extremal Problems*, Technical report TR–RI 94–152, University of Paderborn, Paderborn, Germany, 1994.

[8] S. L. Bezrukov, *Edge isoperimetric problems on graphs*, in Graph Theory and Combinatorial Biology, Bolyai Soc. Math. Stud. 7, L. Lovasz, A. Gyarfas, G. O. H. Katona, A. Recski, and L. Szekely, eds., János Bolyai Math. Soc., Budapest, 1999, pp. 157–197.

[9] B. Bollobás, *Combinatorics*, Cambridge University Press, Cambridge, UK, 1986.

[10] B. Bollobás and I. Leader, *An isoperimetric inequality on the discrete torus*, SIAM J. Discrete Math., 3 (1990), pp. 32–37.

[11] B. Bollobás and I. Leader, *Edge-isoperimetric inequalities in the grid*, Combinatorica, 11 (1991), pp. 299–314.

[12] F. R. K. Chung, *Spectral Graph Theory*, CBMS Reg. Conf. Ser. Math. 92, AMS, Providence, RI, 1997.

[13] F. R. K. Chung and P. Tetali, *Isoperimetric inequalities for Cartesian products of graphs*, Combin. Probab. Comput., 7 (1998), pp. 141–148.

[14] L. H. Harper, *Optimal assignment of numbers to vertices*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 131–135.

[15] L. H. Harper, *On an isoperimetric problem for Hamming graphs*, Discrete Appl. Math., 95 (1999), pp. 285–309.

[16] D. J. Kleitman, M. M. Krieger, and B. L. Rothschild, *Configurations maximizing the number of pairs of Hamming–adjacent lattice points*, Stud. Appl. Math., 50 (1971), pp. 115–119.

[17] I. Leader, *private communication*, 2000.

[18] F. T. Leighton, *Introduction to Parallel Algorithms and Architectures: Arrays · Trees · Hypercubes*, Morgan Kaufmann, San Mateo, CA, 1992.

[19] J. H. Lindsey, II, *Assignment of numbers to vertices*, Amer. Math. Monthly, 71 (1964), pp. 508–516.

[20] I. G. Macdonald, *Symmetric Functions and Hall Polynomials*, 2nd ed., Clarendon Press, Oxford, 1995.

[21] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, San Diego, 1979.

[22] B. Mohar, *Isoperimetric numbers of graphs*, J. Combin. Theory Ser. B, 47 (1989), pp. 274–291.

[23] K. Nakano, *Linear layouts of generalized hypercubes*, in Graph-Theoretic Concepts in Computer Science, Lecture Notes in Comput. Sci. 790, J. van Leeuwen, ed., Springer-Verlag, Berlin, 1994, pp. 364–375.

[24] O. Riordan, *An ordering on the even discrete torus*, SIAM J. Discrete Math., 11 (1998), pp. 110–127.

[25] D.-L. Wang and P. Wang, *Discrete isoperimetric problems*, SIAM J. Appl. Math., 32 (1977), pp. 860–870.

SIAM J. DISCRETE MATH.
Vol. 17, No. 2, pp. 237–248

A 5/8 APPROXIMATION ALGORITHM FOR THE MAXIMUM ASYMMETRIC TSP[*]

MOSHE LEWENSTEIN[†] AND MAXIM SVIRIDENKO[‡]

Abstract. The maximum asymmetric traveling salesperson problem, also known as the taxicab rip-off problem, is the problem of finding a maximally weighted tour in a complete asymmetric graph with nonnegative weights.

We propose a polynomial time approximation algorithm for the problem with a 5/8 approximation guarantee. This (1) improves upon the approximation factors of previous results and (2) presents a simpler solution to the previously fairly involved algorithms. Our solution uses a simple linear programming formulation. Previous solutions were combinatorial. We make use of the linear programming in a novel manner and strengthen the path-coloring method originally proposed in [S. R. Kosaraju, J. K. Park, and C. Stein, *Long tours and short superstrings*, in Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science, 1994, pp. 166–177].

Key words. approximation algorithms, traveling salesperson, linear programming, graph theory

AMS subject classifications. 68W25, 68W40, 68R05, 68R10

DOI. 10.1137/S0895480102402861

1. Introduction. Problem formulation.

In the maximum asymmetric traveling salesman problem, max TSP for short, we are given a complete weighted directed graph $G = (V, E, w)$ with nonnegative edge weights $w_{uv} \geq 0, (u, v) \in E$, and need to find a closed tour of maximum weight visiting all vertices exactly once.

Motivation. The minimization variant of this problem is one of the most studied and well-known optimization problems and has many applications [12]. The Max TSP is also a well-studied and well-motivated optimization problem. We shall shortly mention several applications.

Since Max TSP is an NP-hard optimization problem, it is desirable to design approximation algorithms for this problem with good performance guarantees. We say that an algorithm for a maximization problem has performance guarantee $\rho \leq 1$ if it always delivers a solution with value at least ρ times the value of the optimal solution. (For a minimization problem $\rho \geq 1$ the algorithm should always deliver a solution with value at most ρ times the optimal value.)

The following applications motivate Max TSP:

1. Shortest superstring problem. Given a collection of strings $s_1, \ldots, s_n$ we seek the shortest possible string S such that every string in the collection is a substring of S, i.e., for all $i, S = S' s_i S''$. This problem arises in DNA sequencing and has applications to data compression as well. Kosaraju, Park, and Stein [13] noted that there is an implicit reduction from shortest superstring to Max TSP in the proof of [5]. An ever tighter reduction is given by Breslauer, Jiang, and Jiang [6]. Specifically, they show that a ρ approximation factor for Max TSP implies a $(3.5 - 1.5) \cdot \rho$ approximation for shortest superstring.

[*] Received by the editors August 7, 2002; accepted for publication (in revised form) September 23, 2002; published electronically November 4, 2003.

http://www.siam.org/journals/sidma/17-2/40286.html

[†] Department of Computer Science, Bar Ilan University, 52900 Ramat Gan, Israel (moshe@cs.biu.ac.il).

[‡] IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 (sviri@us.ibm.com).

2. Maximal compression problem. This problem is the sibling of the shortest superstring problem. Given a collection of strings $s_1, \ldots, s_n$ we seek a string S such that every string in the collection (1) is a substring of S and (2) maximizes $\sum_i |s_i| - |S|$. The shortest superstring's optimal solution is equal to the optimal solution of this problem, but the approximate solutions can differ greatly approximationwise. This problem, which arises in various data compression problems, was analyzed by Tarhio and Ukkonen [18] and by Turner [19]. In this setting the vertices represent strings and an edge between two "strings" is weighted with the amount of maximum overlap between these strings. The optimal compression is equivalent to the weight of the maximal Hamiltonian path. With the creation of a special vertex representing the start and end of the Hamiltonian cycle, the maximal compression problem is equivalent to Max TSP on this graph. Hence a ρ approximation for Max TSP implies a ρ approximation for the maximal compression problem.
3. Minimum asymmetric {1,2}-TSP. Introduced by Papadimitriou and Yannakakis [16], this problem is that of the classical minimum TSP in the asymmetric case when the edge weights are 1 or 2. The problem can easily be transformed into the maximum variant, where the weights of 2 are replaced by 0. It is easily verifiable that a ρ factor approximation for Max TSP implies a $2 - \rho$ approximation algorithm for minimum asymmetric {1,2}-TSP.

Previous results. As we already noted, the Max TSP is an NP-hard optimization problem. Moreover, Papadimitriou and Yannakakis [16] proved that this problem is Max SNP-hard and, therefore, we cannot obtain a polynomial time approximation scheme for this problem unless $P = NP$. This result was refined by Engebretsen [7]; his result implies that there is no $(\frac{2803}{2804} + \varepsilon)$ approximation algorithm for Max TSP for any $\varepsilon > 0$ unless $P = NP$ (note that he proved this result for minimum asymmetric $\{1, 2\}$-TSP). Recently, Engebretsen and Karpinsky [8] improved the negative result to $319/320 + \varepsilon$. The first polynomial time algorithm with proven performance guarantee for this problem is due to Fisher, Nemhauser, and Wolsey [10]. They noticed that the algorithm which finds a minimum cycle cover on the input graph, deletes the cheapest edge in each cycle, and patches all cycles together has performance guarantee 1/2. Other approximation algorithms which improve the approximation factors appear in [14, 13, 2]. The best known result to date was an 8/13 factor [2].

In the case when all weights are either 0 or 1, Vishwanathan [20] obtained a 7/12 approximation and Bläser and Siebert [4] obtained a 2/3 approximation algorithm exploiting an approach due to Papadimitriou and Yannakakis [16]. Many other variants of approximation algorithms for Max TSP appear in the literature. A good survey on Max TSP, also containing many results that appeared in the Russian literature, is [1].

Our results. In this paper we present a new algorithm which achieves a 5/8 approximation factor for Max TSP. The algorithm uses a simple LP formulation of a cycle cover which does not contain 2-cycles. This is a very weak variant of the full classical *subtour elimination* constraints. We solve this LP and transform it into a collection of cycle covers with useful properties. We then remove edges to satisfy a generalized version of the path-coloring lemma due to Kosaraju, Park, and Stein [13]. This yields many Hamiltonian cycles, of which one must contain weight at least 5/8 of the maximal weight Hamiltonian cycle.

Our result immediately implies a better result for the maximal compression problem. It also implies an approximation algorithm for the minimum $\{1, 2\}$-TSP with performance guarantee 11/8. Recently, a better performance guarantee of 4/3 was shown for this special case [4]. Another implication is a simple $(41/16 = 2.5 + 1/16)$ approximation algorithm for the shortest superstring problem. The best known approximation algorithm for this problem has performance guarantee 2.5 [17], yet is quite complicated.

2. Algorithm and analysis. In this section we present our 5/8 approximation algorithm. We will assume that the input graph G has an even number of vertices. For the case of an odd number of vertices we show a reduction to the even number case in section 3. This assumption is made in order to be able to use perfect matchings in our algorithm (more specifically, maximum matchings in a complete graph).

We begin with some preliminary notation. For a vertex set V let $K(V)$ denote the edge set $K(V) = (V \times V) \setminus \{(v, v) \mid v \in V\}$. We will consider weighted directed graphs $G = (V, K(V), w)$, where $w : K(V) \to \mathcal{Q}_{\geq 0}$. With w_{uv} we shorthand $w(u, v)$. The optimal solution to the Max TSP problem will be denoted *opt*.

A *cycle cover* C for a graph is a set of cycles such that each vertex participates in exactly one cycle. The weight of a cycle cover, denoted $w(C)$, is the sum of the weights of its edges. A directed multigraph is said to be *d-regular* if the indegree and the outdegree of each vertex is equal to d.

A set of edges E is called *k-path-colorable* if the set of edges can be partitioned into k collections of vertex disjoint paths. The collections $\langle E_1, \dots, E_k \rangle$ are said to *color-partition* E. Let E be a set of edges that is 1-path-colorable. We say that an edge $e \notin E$ is *path-compatible* with E if $\{e\} \cup E$ is 1-path-colorable.

2.1. Generic algorithm and previous solutions. The following generic algorithm first appeared in [10] and was also used in [13, 2, 20].

Generic Algorithm.
Step 1. Compute a maximum-weight cycle cover C for G.
Step 2. Use C to obtain a set P of vertex-disjoint paths.
Step 3. Construct a tour by patching together the paths of P.

In [10] the algorithm uses a simple method. To obtain the set P in step 2 the lightest edge in each cycle was discarded, yielding a 1/2 approximation factor. In [13, 2, 20] better approximations were achieved by slightly adapting the graph and finding a maximum matching M along with the maximum cycle cover. The main idea is to discard some fraction of the edges to obtain a 2-path-colorable set of edges.

If $C \cup M$ was 2-path-colorable with color-partition $\langle E_1, E_2 \rangle$, then choosing the set of edges with the larger weight would yield a solution $\geq \frac{3}{4}$ *w(opt)* since $w(M) + w(C) \geq 1.5w(opt)$.

The algorithms will focus on the fraction of edges needed to be discarded. The results in [13, 2] are achieved by a trade-off between several applications (three and four, respectively) of the algorithms with different discarding schemes. All discarding schemes rely on the following graph coloring property.

Lemma 2.1 (KPS lemma; see [13, 3]). *Let G be a directed multigraph such that* (1) *each vertex has indegree at most* 2, *outdegree at most* 2, *and total degree at most* 3, *and* (2) *the graph does not contain any* 2*-cycles. Then the edges of G are* 2*-path-colorable.*

It needs to be pointed out that in [13] only the formulation of the Lemma is contained. A partial proof appears in [2] and a full proof appears in [3].

For our solution we will need a more general version of the KPS lemma allowing us to color certain graphs with 2-cycles. The 2-cycles that remain undesirable are those that are on a cycle other than this 2-cycle. Namely, for a cycle $\langle v_1, \ldots, v_k, v_1\rangle$ in a directed multigraph G a *back-edge* is an edge (v_i, v_{i-1}), where $i-1$ is k if i is 1. Note that for the case $k = 2$, (v_2, v_1) of the cycle is not called a back-edge, yet a different (v_2, v_1) (which can happen in a multigraph) is called a back-edge; see the left-hand figure in Figure 1. The following lemma is the desired generalization.

LEMMA 2.2 (path-coloring lemma). *Let $G = (V, E)$ be a directed multigraph such that* (1) *each vertex has indegree at most* 2, *outdegree at most* 2, *and total degree at most* 3, *and* (2) *there are no back-edges on any cycle in G. Then E is* 2*-path-colorable.*

Proof. The lemma will follow from the KPS lemma if we eliminate all 2-cycles from G. To this end one 2-cycle will be eliminated while maintaining the properties of the lemma, not creating any new 2-cycles, and ensuring that the eliminated 2-cycle can be 2-path-colored in compatibility with the rest of the graph. Hence, one can eliminate all 2-cycles and the result will follow.

Let $(u, v), (v, u) \in E$ be a 2-cycle in G. Since total degree of u is at most 3 there is at most one other edge, e_1, incident on u, either (x, u) or (u, x) for some vertex x. Vertex x cannot be equal to v because of property (2). Likewise, there is at most one edge, e_2, (v, w) or (w, v). There are four cases.

Case 1. u and v both have total degree 2: this 2-cycle is isolated and is 2-path-colorable independent of $G \setminus \{u, v\}$.

Case 2. Only one of u and v has total degree 3. This case is a special case of case 4.

In the next two cases both u and v have total degree 3.

Case 3. One of $\{(u, v), (v, u)\}$ is path-compatible with both e_1 and e_2. Assume w.l.o.g. that $e_1 = (x, u)$ and that $e_2 = (v, w)$. Set $V' = V \setminus \{u, v\}$. The subgraph induced by V', G' has exactly four edges fewer than G. Add edge (x, w) to G'. It is straightforward to check that the properties are maintained. Moreover, by property (2), no new 2-cycle is created. For a 2-path-coloring of G, let $\langle E_1, E_2\rangle$ be a 2-path-coloring of G' such that $(x, w) \in E_1$. Then $\langle E_1 \cup \{(x, u), (u, v), (v, w)\} \setminus \{x, w\}, E_2 \cup \{(v, u)\}\rangle$ can be easily verified to be a 2-path-coloring of G.

Case 4. Each of (u, v) and (v, u) is path-compatible with exactly one of e_1 and e_2. Contract u and v into one vertex z discarding edges (u, v) and (v, u). All properties are maintained and a new 2-cycle is not created since z has either indegree 0 or outdegree 0. A 2-path-coloring in the contracted graph can be extended by assigning (u, v) and (v, u) the color of their path-compatible edge e_1 or e_2. □

2.2. LP for cycle cover and algorithm outline. In our solution we maintain the generic structure of the algorithm but propose a novel way of obtaining the vertex-disjoint paths. Rather than finding a cycle cover using combinatorial methods we formulate an LP for cycle cover and add an additional constraint. In the integer programming corresponding to this LP the constraint forbids 2-cycles.

This LP is then solved and used to construct a collection of cycle covers. To each cycle cover in the collection we add a maximum matching, as in [13, 2]. We then transform this collection of graphs, by discarding and moving edges, into a collection of graphs that are 2-path-colorable. The collection of vertex-disjoint paths with largest weight is chosen from amongst all graphs. Our discarding scheme is different from previous schemes; in particular, we will be transferring edges between graphs in the collection.

The LP is formulated as following.

LP FOR CYCLE COVER WITH TWO-CYCLE CONSTRAINT.

Max $\sum_{(u,v)\in K(V)} w_{uv}x_{uv}$ subject to

$\sum_u x_{uv} = 1$	for all v	(indegree constraints)
$\sum_v x_{uv} = 1$	for all u	(outdegree constraints)
$x_{uv} + x_{vu} \leq 1$	for all $u \neq v$	(two-cycle constraints)
$x_{uv} \geq 0$	for all $u \neq v$	(nonnegativity constraints)

Let $\{x^*_{uv}\}_{uv\in K(V)}$ be a solution for the LP. Set D to be the minimal integer such that for all $(u,v) \in K(V)$, $D \cdot x^*_{uv}$ is integral. Denote by $k \cdot (u,v)$ the multiset containing k copies of the edge (u,v). Define the weighted multigraph $D \cdot G = (V, \hat{E}, w)$, where $\hat{E} = \{(D \cdot x^*_{uv}) \cdot (u,v) \mid (u,v) \in K(V)\}$. Note that D may be exponential in the graph size which creates a problem. In this section we assume that D is polynomial in the size of the graph and show a polynomial time implementation for general D in section 4.

Note that it follows from the degree constraints of the LP that $D \cdot G$ is D-regular. Hence, we can apply the following lemma to the multigraph $D \cdot G$.

LEMMA 2.3. 1. *Let G be a d-regular multigraph. The edges of G can be partitioned into d cycle covers.*

2. *If the indegree and the outdegree of any vertex in the multigraph G is at most d, then edges of G can be partitioned into d collections of vertex-disjoint cycles and paths.*

This lemma is well known and is a straightforward consequence of König's theorem on the edge colorings of bipartite multigraphs [9].

Another ingredient necessary for our algorithm is a maximum matching in a directed graph. To find a maximum matching, the directed graph can be converted into an undirected graph by setting the weight of edge (u,v) to be the maximum of $\{w_{uv}, w_{vu}\}$. A maximum matching in the directed graph now has a one-to-one correspondence with one in the undirected graph.

We now present the algorithm.

MAX TSP ALGORITHM.
Step 1. *Solve the LP.*
Step 2. *Set G' to be $D \cdot G$.*
Step 3. *Split G' into D cycle covers $C_1, \ldots, C_D$.*
Step 4. *Find a maximum matching M in G.*
Step 5. *Transform $C_1 \cup M, \ldots, C_D \cup M$ into D graphs $G_1, \ldots, G_D$ such that each G_i is* 2*-path-colorable by Lemma* 2.2 *with color-partition $\langle P_i, P'_i \rangle$.*
Step 6. *Choose the collection with largest weight from collections $\{P_1, P'_1, \ldots, P_D, P'_D\}$.*
Step 7. *Patch the chosen collection to a Hamiltonian tour.*

2.3. Transformation into D 2-path-colorable graphs.

Define the graphs G_i to be $C_i \cup M$. The graphs need to be transformed so that each will be 2-path-colorable. To achieve this we will be moving edges between the different graphs and removing edges when necessary. Our goal is to do this in a manner that will allow us to bound the fraction of the overall weight that is removed when edges are deleted.

In this section we will do this in such a manner that the overall weight of the removed edges is at most half the weight of M on average for each of the graphs G_i. So, on average for each G_i, the weight will be bounded from below by $w(C) + \frac{1}{2}w(M) \geq w(opt) + \frac{1}{4}w(opt) = \frac{5}{4}w(opt)$. Hence taking the graph with heaviest weight and coloring the edges with a 2-path-coloring yields a 5/8 approximation for Max TSP.

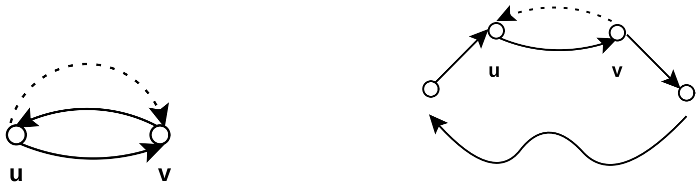


FIG. 1. *Dashed edges are from M and solid edges from the cycle cover. On the left, configuration* 1, *an $\mathcal{F}_1$, and, on the right, configuration* 2.

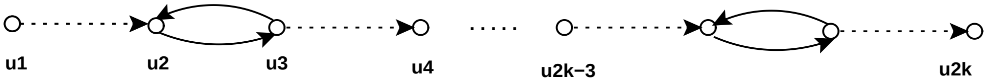


FIG. 2. *Configuration* 3*—alternating edges from M and cycles from the cycle cover. Possibly,* $(u_{2k-1}, u_{2k}) = (u_1, u_2)$.

At the start each G_i contains a cycle cover and a matching. Therefore property (1) of the path-coloring lemma is satisfied. The transformation which we will shortly describe needs to ensure that there are no back-edges on any cycle, i.e., property (2) of the path-coloring lemma. There are three possible configurations where G_i might contain a cycle, not necessarily from the cycle cover, with a back-edge upon it.

Configuration 1. A 2-cycle from C_i with a parallel edge from the matching; see $\mathcal{F}_1$ in Figure 1.

Configuration 2. An edge (u, v) is on some cycle in C_i of size ≥ 3 and there is an edge $m = (v, u) \in M$; see Figure 1.

Configuration 3. A 2-cycle $\{(u, v), (v, u)\}$ appears in C_i, where, say, (v, u) is the back-edge. In this case, since (u, v) is on a "cycle," there must be edges (x, u) and (v, w) from the matching M, where $x \neq w$ and $x, w \notin \{u, v\}$. In fact, there may be a chain of interchanging matching edges and 2-cycles; see Figure 2. It is important to note that to comply with property (2) of the path-coloring lemma it is sufficient to break chains at any location on the chain.

Formally, $u_1, u_2, \ldots, u_{2k}$ is a *chain* in G_i if (1) $(u_{2j-1}, u_{2j}) \in M$ for $1 \leq j \leq k$ and (2) $(u_{2j}, u_{2j+1}), (u_{2j+1}, u_{2j}) \in C_i$ for $1 \leq j \leq k$, and (3) $k \geq 2$. A *maximal chain* in G_i is a chain that cannot be extended to a larger chain.

It is quite easy to verify, using a case-based analysis of a pair $\{u, v\}$, that these three bad configurations are the only ones violating property (2) of the path-coloring lemma in a graph that is the union of a matching and a cycle cover.

2.3.1. Removing Configuration 1 and its counterpart Configuration 3.

To remove the undesired configurations we will need to delete some of the edges from the graphs. However, to obtain a lower bound on the fraction of edges we remove, we prove the following.

LEMMA 2.4 (balancing lemma). *Let $m = (u, v) \in M$. If there are f graphs in which m is contained in an $\mathcal{F}_1$, then there are at least f other graphs in which the cycle cover does not contain both (u, v) and (v, u).*

Proof. By the 2-cycle constraint, $x^*_{uv} + x^*_{vu} \leq 1$. Hence $D \cdot (x^*_{uv} + x^*_{vu}) \leq D$. Since the edge (u, v) appears in $D \cdot x^*_{uv}$ covers, and the edge (v, u) appears in $D \cdot x^*_{vu}$ covers the number of edges (u, v) and (v, u) in all cycle covers combined is $\leq D$.

By definition, for each G_i containing an $\mathcal{F}_1$ on vertices $\{u, v\}$ both (u, v) and (v, u) appear in C_i. If there are f graphs containing an $\mathcal{F}_1$ on vertices $\{u, v\}$, then these f graphs contain $2f$ cycle cover edges between u and v. Hence, there are at most $D - 2f$ *other* graphs G_j containing exactly one cycle cover edge (u, v) or (v, u). Since there are D graphs overall, f containing an $\mathcal{F}_1$ on vertices $\{u, v\}$ and at most $D - 2f$ graphs containing at least one edge from a cycle cover between u and v, it follows that there are at least f graphs containing no edge from a cycle cover between u and v. □

Consider an edge $m = (u, v) \in M$. m appears in each of the D graphs G_i. It follows from the balancing lemma that each m which appears in an $\mathcal{F}_1$, say in G_i, can be paired uniquely with an m, say in G_j, where (u, v) and (v, u) do not appear in C_j. An *m-pair* is a triplet $\langle m, i, j\rangle$, where $m \in M$ is on an $\mathcal{F}_1$ in G_i and its pair is in G_j.

We discard the edge m from the $\mathcal{F}_1$ in G_i, thus eliminating Configuration 1. If we can guarantee that its pair, the m in G_j, is not removed, then we remove only one out of two edges to eliminate Configuration 1. However, since the m in G_j may participate in Configuration 3, we need a slightly more sophisticated method which eliminates both undesired configurations together. Note that the m in G_j cannot be of type Configuration 2.

Prior to removing the edges of M that appear in $\mathcal{F}_1$, we identify all maximal chains, i.e., Configuration 3, in all the graphs. Remember, breaking a chain at any point eliminates the problem of violating property (2) of the path-coloring lemma. One way to break the chain is to remove an edge on the chain. Another possibility is to reverse one of the edges from M on the chain. One must be careful not to reverse all edges from M on the chain, because otherwise a reverse cycle might be created once again violating property (2) of the path-coloring lemma. Yet, reversing at least one edge and making sure that at least one edge does not get reversed is sufficient to "break" the chain.

Consider a pair $\langle m, i, j\rangle$ where m is on such a maximal chain in C_j. We check whether any edge has been reversed on this maximal chain. If one has, then there is no Configuration 3 "problem" in G_j and we simply remove m from G_i to eliminate the Configuration 1 from G_i. If no edge has been reversed on the maximal chain, then we reverse it, i.e., $m = (u, v)$ is removed from G_j and (v, u) is added to G_j. In parallel, we remove (v, u) from G_i. Hence overall we have removed one copy of m. Both configurations have been fixed.

2.3.2. Removing Configuration 3. The only type of Configuration 3 necessary to fix, after removing edges of type Configuration 1, are maximal chains that have not been paired with Configuration 1 edges. Removing the cheapest edge from M appearing on a chain breaks the chain as desired. Moreover, the weight removed is less than half the weight of the edges from M on this maximal chain, upper bounding the removal cost.

2.3.3. Removing Configuration 2. For those back-edges violating property (2) of the path-coloring lemma because of Configuration 2, the following lemma provides a removal scheme. The scheme removes at most 1/2 of the weight of the violating back-edge.

LEMMA 2.5. *Let $P = (x, u), (u, v), (v, w)$ be a portion of a cycle from C_i in G_i and let $(v, u) \in M$. Let W be the overall weight of the edges in G_i. Then by* (1) *removing (u, v) or* (2) *removing (v, u) and adding another (u, v), the overall weight will be at least $W - \frac{1}{2}w_{vu}$.*

Proof. If $w_{uv} \leq \frac{1}{2} w_{vu}$, then (1) removing edge (u, v) gives the desired result. Otherwise, $w_{uv} \geq \frac{1}{2} w_{vu}$ and (2) yields the result. □

Both (1) and (2) remove the undesired 2-cycle. Hence, this fixes the violation of property (2) in the path-coloring lemma.

2.4. Putting it all together. Step 5 of our algorithm works as follows.

Transformation to back-edge–free cycles.
Step 5a. Identify edges $m = (u, v) \in M$ that appear on an $\mathcal{F}_1$ in each of the graphs.
Step 5b. Pair, uniquely, each such edge m in G_i with an m in some G_j such that $(u, v), (v, u) \notin C_j$.
Step 5c. Identify all maximal chains of Configuration 3.
Step 5d(a). For each m on an $\mathcal{F}_1$, if its pair is not on a maximal chain, remove it.
Step 5d(b). If its pair is on a maximal chain, then if $\exists e \in M$ on the maximal chain such that e has been reversed, remove m from the $\mathcal{F}_1$.
Step 5d(c). Otherwise, reverse m on the maximal chain, i.e., remove $m = (u, v)$ from G_j and replace it with (v, u). On G_i remove (v, u).
Step 5e. For each maximal chain where no edge has been reversed, delete the cheapest edge from M on the chain.
Step 5f. To each appearance of $m = (u, v) \in M$ in each graph G_i such that (v, u) appears in C_i but (u, v) does not, apply Lemma 2.5.

Theorem 2.6. *The algorithm produces a Hamiltonian path of weight* $\geq \frac{5}{8} w(opt)$, *where* $w(opt)$ *is the weight of Max TSP.*

Proof. The Hamiltonian cycle in G corresponds to a feasible solution of the LP. The optimal solution to the LP is $\{x^*_{uv}\}_{(u,v) \in K(V)}$. Therefore, $w(opt) \leq \sum_{(u,v) \in K(V)} w_{uv} x^*_{uv}$. Hence $D \cdot w(opt) \leq D \cdot \sum_{u,v \in K(V)} w_{uv} x^*_{uv}$. It follows from Lemma 2.3 that after Step 3, $\sum_{i=1}^{D} w(C_i) = D \cdot \sum_{u,v \in K(V)} w_{uv} x^*_{uv}$. Obviously, $w(M) \geq \frac{1}{2} \cdot w(opt)$. Since M was added to each of the graphs G_i, the weight of the graphs before the transformation was $\sum_{i=1}^{D} w(C_i) + D \cdot w(M)$. From the discussion in section 2.3 the overall weight deleted from all the graphs together is $\leq \frac{1}{2} \cdot D \cdot w(M)$. Hence the overall weight of the graphs after Step 5 is at least $\sum_{i=1}^{D} w(C_i) + \frac{1}{2} \cdot D \cdot w(M) = D \cdot \sum_{u,v \in K(V)} w_{uv} x^*_{uv} + \frac{1}{2} \cdot D \cdot w(M) \geq D \cdot w(opt) + \frac{1}{4} \cdot D \cdot w(opt) = \frac{5}{4} \cdot D \cdot w(opt)$. Since each of the D graphs is 2-path-colorable, it follows that there are $2D$ path collections to choose from. Choosing the heaviest yields one of weight $\frac{\frac{5}{4} \cdot D \cdot w(opt)}{2D} = \frac{5}{8} \cdot w(opt)$. Completing the path collection to a Hamiltonian tour cannot decrease the weight. □

3. Algorithm for graphs with odd number of vertices. If the input graph G has an odd number of vertices, then the maximum matching M in G does not necessarily satisfy the inequality $w(M) \geq w(opt)/2$. In this case we show how to reduce the problem to the problem on a graph with an even number of vertices. Assume that we have guessed two consecutive edges $(v_1, v_2), (v_2, v_3)$ of some optimal tour for Max TSP in G; this is done by enumerating over all directed paths of length 2. Find the maximum matching M_1 in the graph $G \setminus \{v_1, v_2, v_3\}$ and maximum matching M_2 in the graph $G \setminus \{v_2\}$. We claim that $w(M_1) + w(M_2) + w_{v_1 v_2} + w_{v_2 v_3} \geq w(opt)$ and therefore $\max\{w(M_1) + w_{v_1 v_2} + w_{v_2 v_3}, w(M_2)\} \geq w(opt)/2$. The reason that $w(M_1) + w(M_2) + w_{v_1 v_2} + w_{v_2 v_3} \geq w(opt)$ is that the optimal tour can be partitioned into two edge-disjoint graphs: one is a matching in $G \setminus \{v_1, v_2, v_3\}$ plus two edges

$(v_1, v_2), (v_2, v_3)$ and another is a matching in $G \setminus \{v_2\}$.

If $w(M_2) \geq w(M_1) + w_{v_1v_2} + w_{v_2v_3}$, then we can apply the previous algorithm without modifications since in this case $w(M_2) \geq w(opt)/2$ and M_2 is a matching.

If $w(M_2) < w(M_1) + w_{v_1v_2} + w_{v_2v_3}$, then we delete vertex v_3 from G and redefine the weight of the edge (v_1, v_3) to be $w_{v_1v_2} + w_{v_2v_3}$. After that we apply our algorithm to the optimal fractional solution of the LP defined on the new instance of the problem and matching $M_1 \cup \{(v_1, v_3)\}$. If P is the heaviest collection of paths delivered by the algorithm, then we add vertex v_2 back to the graph G. If P contains edge (v_1, v_3), then we delete it and add two edges (v_1, v_2) and (v_2, v_3) instead. Since the optimal value of the LP on the new instance of the problem is an upper bound on the $w(opt)$, we obtain that $w(P) \geq w(opt)/2 + (w(M_1) + w_{v_1v_2} + w_{v_2v_3})/4 \geq 5/8w(opt)$.

4. Polynomial time implementation of the algorithm. The algorithm described in the previous section might obviously run in exponential time. The reason is that the number D (defined as the minimum number such that all numbers Dx^*_{uv} are integers) can be exponentially big and therefore the algorithm will produce an exponential number of cycle covers. The simple way to deal with this problem is to round each number x^*_{uv} down to the nearest multiple of ε/n^2, where $\varepsilon > 0$ is some precision parameter such that $\varepsilon = 1/s$ for some integer s. Let $\bar{x}_{uv}$ be the value of variable x_{uv} obtained after rounding x^*_{uv}. Let $\bar{D}$ be a minimum number such that all numbers $\bar{D}\bar{x}_{uv}$ are integers. Clearly, $\bar{D} \leq sn^2$. Consider the multigraph $\bar{D} \cdot G$, i.e., the multigraph having $\bar{D}\bar{x}_{uv}$ copies of edge (u, v). Since we rounded x^*_{ij} down, the indegree and outdegree of any vertex in $\bar{D} \cdot G$ are at most $\bar{D}$; moreover, there are at most $\bar{D}$ edges between any pair of vertices u and v by the 2-cycle constraint. By applying Lemma 2.3 we can partition $\bar{D} \cdot G$ into $\bar{D}$ collections of vertex-disjoint cycles and paths. By adding the matching M to each of these collections and eliminating all three types of "bad" configurations as in the previous section, we obtain $\bar{D}$ graphs satisfying conditions of Lemma 2.2. Therefore, applying Lemma 2.2 to each of these graphs we get $2\bar{D}$ partial tours which can be completed to the Hamiltonian cycle. We now estimate the value of the best Hamiltonian path obtained this way. If W is a maximum weight of the edge in the input graph G, then clearly $W \leq w(opt)$. The total weight of the graph $\bar{D} \cdot G$ is

$$\bar{D} \sum_{u,v \in V} w_{uv}\bar{x}_{uv} \geq \bar{D} \sum_{u,v \in V} w_{uv}(x^*_{uv} - \varepsilon/n^2) \geq \bar{D} \sum_{u,v \in V} w_{uv}x^*_{uv} - \bar{D}\varepsilon W.$$

Therefore, the Hamiltonian cycle delivered by the algorithm has weight at least $\frac{5}{8}w(opt) - \frac{\varepsilon W}{2} \geq (\frac{5}{8} - \frac{\varepsilon}{2})w(opt)$.

We now describe a more complicated procedure of implementing our algorithm in polynomial time without losing ε in the performance guarantee. A matrix is called *doubly stochastic* if all of its entries are nonnegative and the sum of all the elements in any row or column is exactly one. A doubly stochastic matrix with integer entries is called a *permutation matrix*.

Lemma 4.1 (Birkhoff–von Neumann [15]). *Any doubly stochastic $n \times n$ matrix can be represented as a convex combination of at most n^2 permutation matrices and such representation can be found in polynomial time.*

This lemma can be derived by recursive application of Hall's theorem [9]. The actual representation can be found by at most n^2 applications of the algorithm for finding a perfect matching in bipartite graph.

The optimal solution of the LP $X = \{x^*_{uv}\}_{u,v \in K(V)}$ is a doubly stochastic matrix by indegree, outdegree, and nonnegativity constraints. Applying Lemma 4.1 we obtain

$X = \sum_{i=1}^{n^2} \lambda_i \Pi_i$, where Π_i are permutation matrices and $\lambda_i \geq 0$ are coefficients in the convex combination, i.e., $\sum_{i=1}^{n^2} \lambda_i = 1$. Each permutation matrix Π_i corresponds to a cycle cover C_i in the original graph G. Moreover, if we consider the multigraph $D \cdot G$, where D is the minimal number such that $\lambda_i D$ are integers for all $i = 1, \ldots, n^2$ (such a D is not necessarily polynomially bounded), then Π_i represents $\lambda_i D$ identical cycle covers in the cycle cover decomposition of $D \cdot G$ guaranteed by Lemma 2.3. We now implement our algorithm working with cycle covers $C_i, i = 1, \ldots, n^2$, instead of working with D cycle covers explicitly.

In the first stage we add the matching M to each of the cycle covers $C_i, i = 1, \ldots, n^2$. If all graphs $G_i = C_i \cup M, i = 1, \ldots, n^2$, don't contain "bad" configurations (Figures 1 and 2), then by the Lemma 2.2 we can find a 2-path-coloring of each $C_i \cup M$. Let P_i and P_i' be collections of paths corresponding to a 2-path-coloring of $C_i \cup M$. Then the weight of the heaviest collection P of paths is at least $3/4w(opt)$ since $w(P) \geq \sum_{i=1}^{n^2} \lambda_i (w(P_i) + w(P_i'))/2 = \sum_{i=1}^{n^2} \lambda_i (w(C_i) + w(M))/2 \geq 3/4w(opt)$. So, what we need to show is that we can remove all "bad" configurations, as before, by deleting at most half of the edges from M on average.

Removing Configuration 1 and its counterpart Configuration 3. We now prove a generalization of the balancing lemma used in the previous section.

LEMMA 4.2 (generalized balancing lemma). *Let $m = (u, v) \in M$ and let S_2 be a subset of the set $\{1, \ldots, n^2\}$ such that m is contained in an $\mathcal{F}_1$ in each graph $C_i \cup M$ for $i \in S_2$. Then there is a set $S_0 \subset \{1, \ldots, n^2\}$ such that* (1) *for each $i \in S_0$ the cycle cover C_i in the graph $G_i = C_i \cup M$ does not contain (u, v) and (v, u) and* (2) $\sum_{i \in S_0} \lambda_i \geq \sum_{i \in S_2} \lambda_i$.

Proof. Let $S_0 \subseteq \{1, \ldots, n^2\}$ be the set of cycle covers which do not contain (u, v) and (v, u) and let $S_1 = \{1, \ldots, n^2\} \setminus (S_2 \cup S_0)$, i.e., S_1 is the set of cycle covers C_i having exactly one edge ((u, v) or (v, u)) between vertices u and v. By the 2-cycle constraint, $x_{uv}^* + x_{vu}^* \leq 1$. Hence, $\sum_{i \in S_2} 2\lambda_i + \sum_{i \in S_1} \lambda_i \leq 1$. However, $\sum_{i \in S_2 \cup S_1 \cup S_0} \lambda_i = 1$ and therefore $\sum_{i \in S_0} \lambda_i \geq \sum_{i \in S_2} \lambda_i$. □

Consider the edge $m = (u, v) \in M$. Let $\bar{S}_0 \subseteq S_0$ be any minimal subset of S_0 such that $\sum_{i \in \bar{S}_0} \lambda_i \geq \sum_{i \in S_2} \lambda_i$. If $\sum_{i \in \bar{S}_0} \lambda_i > \sum_{i \in S_2} \lambda_i$, then we take any $t \in \bar{S}_0$ and define two copies t' and t'' of the cycle cover C_t having weights in a convex combination $\lambda_{t''} = \sum_{i \in \bar{S}_0} \lambda_i - \sum_{i \in S_2} \lambda_i$ and $\lambda_{t'} = \lambda_t - \lambda_{t''}$. Redefine $\bar{S}_0 = \bar{S}_0 \cup \{t'\} \setminus \{t\}$ and $S_0 = S_0 \cup \{t', t''\} \setminus \{t\}$. By construction we have $\sum_{i \in \bar{S}_0} \lambda_i = \sum_{i \in S_2} \lambda_i$. We now partition the set $\bar{S}_0$ into two sets S_0' and S_0''. The set S_0' consists of those graphs G_i from $\bar{S}_0$ where edge m is on a maximal chain (Configuration 3), and there were no edges reversed on this maximal chain on prior steps. Let $S_0'' = \bar{S}_0 \setminus S_0'$. After that we partition the set S_2 into two sets S_2' and S_2'' in such a way that $\sum_{i \in S_0'} \lambda_i = \sum_{i \in S_2'} \lambda_i$ and $\sum_{i \in S_0''} \lambda_i = \sum_{i \in S_2''} \lambda_i$. If there is no such S_2', then we make two copies j' and j'' of some graph $G_j, j \in S_2$, and define weights $\lambda_{j'}$ and $\lambda_{j''}$ such that $\lambda_j = \lambda_{j'} + \lambda_{j''}$. Put one copy into S_2' and another into S_2''. We pair set S_0' with set S_2' and set S_0'' with set S_2''. For the first pair of sets we reverse edge (u, v) in graphs $G_i, i \in S_0'$, and delete (v, u) in S_2' as we did in section 2.3. For the second pair we just delete edge (u, v) from all $G_i, i \in S_2''$. We repeat this step for all other edges $(u, v) \in M$ which belong to some "bad" configuration $\mathcal{F}_1$. Each such step increases the number of cycle covers in the convex combination by at most two; moreover, we delete at most half of any matching edge on average. Therefore, in the end of this exclusion step we don't have configurations $\mathcal{F}_1$ and we have at most $3n^2$ cycle covers in a convex combination and corresponding graphs G_i.

Removing Configuration 3. We do exactly what we did in section 2.3. We just remove the cheapest edge in any maximal chain that has not been paired with Configuration 1 edges.

Removing Configuration 2. Applying Lemma 2.5, we remove at most half of any matching edge participating in such a configuration.

5. Conclusion.

An obvious open question is to close the gap between the best known positive result 5/8 and negative result 319/320 for Max TSP. On one hand, we don't even know the gap between the value of an optimal fractional solution of the linear programming relaxation considered in this paper and the optimal solution of Max TSP. The worst instance we know has gap 2/3. On the other hand, our LP is rather weak; we could add many valid inequalities known for TSP [11]. The most promising set of constraints is

$$\sum_{u,v\in S} x_{uv} \leq |S| - 1 \text{ for all } S \subset V.$$

These constraints are known as the *subtour elimination* inequalities and they guarantee that the graph $D \cdot G$ considered in section 2 is strongly D-edge-connected. Unfortunately, we don't know how to use this property in the rounding procedure. However, we point out that it does eliminate the example showing a 2/3 gap for the LP with weaker constraints.

Acknowledgments. The authors would like to thank A. Ageev, R. Hassin, and A. Yeo for helpful discussions on the subject of the paper.

REFERENCES

[1] A. I. Barvinok, E. Kh. Gimadi, and A. I. Serdyukov, *The maximum Traveling Salesman Problem*, in The Traveling Salesman Problem and Its Variations, G. Gutin and A. Punnan, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 585–607.

[2] M. Bläser, *An $\frac{8}{13}$-approximation algorithm for the asymmetric maximum TSP*, in Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2002, pp. 64–73.

[3] M. Bläser, *An $\frac{8}{13}$-approximation Algorithm for the Asymmetric Max-TSP*, Technical report SIIM-TR-A-01-21, Institute für Informatik und Mathematik, Universtität Lübeck, Germany, 2001.

[4] M. Bläser and B. Siebert, *Computing cycle covers without short cycles*, in Algorithms—ESA 2001, Lecture Notes in Comput. Sci., Springer-Verlag, Berlin, 2001, pp. 368–380.

[5] A. Blum, T. Jiang, M. Li, J. Tromp, and M. Yannakakis, *Linear approximation of shortest superstring*, J. ACM, 31 (1994), pp. 630–647.

[6] D. Breslauer, T. Jiang, and Z. Jiang, *Rotations of periodic strings and short superstrings*, J. Algorithms, 24 (1997), pp. 340–353.

[7] L. Engebretsen, *An explicit lower bound for TSP with distances one and two.* in Lecture Notes in Comput. Sci. 1563, Springer, Berlin, 1999, pp. 373–382.

[8] L. Engebretsen and M. Karpinski, *Approximation hardness of TSP with bounded metrics*, in Proceedings of the 28th International Colloquium on Automata, Languages and Programming, Lecture Notes in Comput. Sci. 2076, F. Orejas, P. G. Spirakis, and J. van Leeuwen, eds., Springer-Verlag, Berlin, 2001, pp. 201–212.

[9] S. Fiorini and R. Wilson, *Edge-Colourings of Graphs*, Research Notes in Mathematics, 16, Pitman, London, 1977.

[10] M. L. Fisher, L. Nemhauser, and L. A. Wolsey, *An analysis of approximations for finding a maximum weight Hamiltonian circuit*, Networks, 12 (1979), pp. 799–809.

[11] M. X. Goemans, *Worst-case comparison of valid inequalities for the TSP*, Math. Programming, 69 (1995), pp. 335–349.

[12] E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. B. Shmoys, eds., *The Traveling Salesman Problem. A Guided Tour of Combinatorial Optimization*, Wiley-Intersci. Ser. Discrete Math., John Wiley, Chichester, UK, 1985.

[13] S. R. Kosaraju, J. K. Park, and C. Stein, *Long tours and short superstrings*, in Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science, 1994, pp. 166–177.

[14] M. M. Koval'ov and V. M. Kotov, *Estimate of the error of series of approximate algorithms*, Vestnik Beloruss. Gos. Univ. Ser. I Fiz. Mat. Mekh., (1986), pp. 44–48 (in Russian).

[15] E. L. Lawler, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, Montreal, London, 1976, p. 189.

[16] C. H. Papadimitriou and M. Yannakakis, *The traveling salesman problem with distances one and two*, Math. Oper. Res., 18 (1993), pp. 1–11.

[17] Z. Sweedyk, $2\frac{1}{2}$*-approximation algorithm for shortest superstring*, SIAM J. Comput., 29 (1999), pp. 954–986.

[18] J. Tarhio and E. Ukkonen, *A greedy approximation algorithm for constructing shortest common superstrings*, Theoret. Comput. Sci., 57 (1988), pp. 131–145.

[19] J. S. Turner, *Approximation algorithms for the shortest common superstring problem*, Inform. Comput., 83 (1989), pp. 1–20.

[20] S. Vishwanathan, *An approximation algorithm for the asymmetric travelling salesman problem with distances one and two*, Inform. Process. Lett., 44 (1992), pp. 297–302.

SIAM J. Discrete Math.
Vol. 17, No. 2, pp. 249–253

CONVEX MATROID OPTIMIZATION[*]

SHMUEL ONN[†]

Abstract. We consider a problem of maximizing convex functionals over matroid bases. It is richly expressive and captures certain quadratic assignment and clustering problems. While generally intractable, we show that it is efficiently solvable when a suitable parameter is restricted.

Key words. combinatorial optimization, convex, optimization, partition, cluster, matroid, greedy algorithm, quadratic assignment, polytope, polynomial time, strongly polynomial time

AMS subject classifications. 05A, 15A, 51M, 52A, 52B, 52C, 68Q, 68R, 68U, 90B, 90C

DOI. 10.1137/S0895480102408559

1. Introduction.

Let $M = (N, \mathcal{B})$ be a matroid over $N := \{1, \ldots, n\}$ with a collection of bases $\mathcal{B} \subseteq 2^N$. Let $w : N \longrightarrow \mathbb{R}^d$ be a weighting of matroid elements by vectors in d-space. For any subset $J \subseteq N$ let $w(J) := \sum_{j \in J} w(j)$ with $w(\phi) := 0$. Finally, let $c : \mathbb{R}^d \longrightarrow \mathbb{R}$ be a convex functional on $\mathbb{R}^d$. We consider the following algorithmic problem.

Convex matroid optimization. Given data as above, find a basis $B \in \mathcal{B}$ maximizing $c(w(B))$.

We begin with some examples of specializations of this problem.

Example 1.1 (linear matroid optimization). This is the special case of our problem with $d = 1$, $w : N \longrightarrow \mathbb{R}$ a weighting of elements by scalars, and $c : \mathbb{R} \longrightarrow \mathbb{R} : x \mapsto x$ the identity. The problem is to find a basis of maximum weight and is quickly solvable by the greedy algorithm.

Example 1.2 (positive semidefinite quadratic assignment). This is the NP-hard problem [8] of finding a vector $x \in \{0,1\}^n$ maximizing $||Wx||^2 = x^T W^T W x$ with W a given $d \times n$ matrix. For fixed d it is solvable in polynomial time [3]. The variant of this problem in which one asks for x with restricted support $|\text{supp}(x)| = r$ is the special case of our problem with $M := U_n^r$ the uniform matroid of rank r over N, with $w(j) := W^j$ the jth column of W for all $j \in N$, and with $c : \mathbb{R}^d \longrightarrow \mathbb{R} : x \mapsto ||x||^2$ the l_2-norm (squared or not). The positive semidefinite quadratic assignment problem can be solved by solving the variant for $r = 0, \ldots, n$ and picking the best x.

Example 1.3 (minimal variance balanced clustering). This is the problem of partitioning a given set $\{w_1, \ldots, w_n\}$ of points in $\mathbb{R}^d$ into two clusters C_1, C_2 of equal size $m := \frac{n}{2}$ so as to minimize the sum of cluster variances given by

$$\frac{1}{m} \sum_{w_j \in C_1} \left|\left| w_j - \left(\frac{1}{m} \sum_{w_j \in C_1} w_j \right) \right|\right|^2 \quad + \quad \frac{1}{m} \sum_{w_j \in C_2} \left|\left| w_j - \left(\frac{1}{m} \sum_{w_j \in C_2} w_j \right) \right|\right|^2 .$$

It can be shown by suitable manipulation of the variance expression that this is the special case of our problem with M the uniform matroid of rank m over N, with

[*] Received by the editors June 3, 2002; accepted for publication (in revised form) July 22, 2003; published electronically November 4, 2003. This research was supported in part by a grant from the Israel Science Foundation (ISF), by a Vice President of the Technion for Research grant, and by the Fund for the Promotion of Research at the Technion.

http://www.siam.org/journals/sidma/17-2/40855.html

[†] The Technion - Israel Institute of Technology, 32000 Haifa, Israel and University of California at Davis, Davis, CA 95616 (onn@ie.technion.ac.il, onn@math.ucdavis.edu, http://ie.technion.ac.il/˜onn).

$w(j) := w_j$ for all $j \in N$, and with $c : \mathbb{R}^d \longrightarrow \mathbb{R} : x \mapsto ||x||^2 + ||w(N) - x||^2$ with $w(N) = \sum_{j=1}^{n} w_j$ the sum of all points.

This problem often arises in the analysis of statistical data and may include constraints requiring some points to be in the same or different clusters, based on some a priori knowledge on the sampled points. Our framework, which allows M to be any matroid, can sometimes accommodate such constraints as well. For instance, if the sample points come in pairs $\{w_{2i-1}, w_{2i}\}$, $i = 1, \ldots, m$, and each cluster is required to contain precisely one point of each pair, then the problem is cast in our framework with the same data as above except that M is now taken to be the transversal matroid with base collection $\mathcal{B} := \{B \subset N : |B \cap \{2i-1, 2i\}| = 1,\ i = 1, \ldots, m\}$.

While the linear matroid optimization problem (Example 1.1) is greedily solvable (cf. [4]), the general convex matroid optimization problem is NP-hard as indicated by Example 1.2. Here, however, we show that for fixed d the problem can be solved efficiently for an arbitrary matroid M and an arbitrary convex functional c. The assumption of fixed d is quite natural in applications: in Example 1.2 it is the rank of the corresponding quadratic form and in Example 1.3 it is the dimension of the sampled points. We assume that c is presented by an *evaluation oracle* that given $x \in \mathbb{R}^d$ returns $c(x)$, and that M is presented by an *independence oracle* that given $J \subseteq N$ asserts whether or not J is an independent set of M. We establish the following theorem.

THEOREM 1.4. *For any fixed d, the convex matroid optimization problem with oracle presented matroid M over $N := \{1, \ldots, n\}$, weighting $w : N \longrightarrow \mathbb{R}^d$, and oracle presented convex functional $c : \mathbb{R}^d \longrightarrow \mathbb{R}$, can be solved in polynomial oracle time using $O(n^{2d-1} \log n)$ operations and queries.*

The computational complexity is measured in terms of the number of real arithmetic operations and oracle queries. For rational input the algorithm is (strongly) polynomial time in the Turing computation model, where the input includes the binary encoding of the weighting $w : N \longrightarrow \mathbb{Q}^d$ and the binary encoding of an upper bound $U := \max_{J \subseteq N} c(w(J))$ on the relevant values of the convex functional, but we do not dwell on the details here.

The special case of the convex matroid optimization problem for *uniform matroids* coincides with the special case of the so-called *shaped partition problem* [10] for *two-parts*. Therefore, the specializations to two-parts of the lower bounds of [1, 2] imply a lower bound of $\Omega(n^{d-1})$ on the complexity of the convex matroid optimization problem. This shows that the linear occurrence of d in the exponent of n in the complexity cannot be avoided, and so our algorithm is optimal in that sense. It would be interesting to determine the best possible constant $1 \leq \alpha \leq 2$ for which the problem is solvable using $O(n^{\alpha d + \beta})$ arithmetic operations and queries. It would also be very interesting to further study a plausible common generalization of the convex matroid optimization problem for arbitrary matroids and the shaped partition problem for arbitrary number of parts.

2. Proof of the theorem.

For a matroid $M = (N, \mathcal{B})$ and a weighting $w : N \longrightarrow \mathbb{R}^d$, consider the following convex polytope:

$$\mathcal{P}_w^M \quad := \quad \operatorname{conv}\{\, w(B) : B \in \mathcal{B}\,\} \quad \subset \quad \mathbb{R}^d.$$

The convex matroid problem can be reduced to maximizing the convex functional c over $\mathcal{P}_w^M$: there will always be an optimal basis $B \in \mathcal{B}$ for which $w(B)$ is a vertex of $\mathcal{P}_w^M$, and so the problem can be solved by picking the best such vertex. However, as the number of matroid bases is typically exponential in n, it is not possible to construct $\mathcal{P}_w^M$ directly in polynomial time. As we shall see, the most efficient way around that is enhanced by constructing the following *zonotope*:

$$\mathcal{P}_w \quad := \quad \sum_{1\le i<j\le n} [-1,1]\cdot(w(i)-w(j)) \quad \subset \quad \mathbb{R}^d.$$

PROPOSITION 2.1. *Fix any d. Then the number of vertices of the zonotope $\mathcal{P}_w$ is $O(\,n^{2(d-1)}\,)$. Further, in polynomial time using that many arithmetic operations, all its vertices can be listed, each vertex v along with a linear functional $a(v)\in\mathbb{R}^d$ maximized over $\mathcal{P}_w$ uniquely at v.*

Proof. It has been known for a long time [9], in the dual setup of hyperplane arrangements, that for fixed d, the number of vertices of any zonotope $\mathcal{Z} := \sum_{i=1}^m[-1,1]\cdot z_i$ generated by m line segments in $\mathbb{R}^d$ is $O(m^{d-1})$. The algorithmic analogue of this result (again in the dual setup of hyperplane arrangements) is provided in [5, 6] (the latter reference provides a necessary correction of the former); it asserts that all vertices of any such zonotope can be enumerated in polynomial time using an optimal number $O(m^{d-1})$ of arithmetic operations, each vertex v along with a linear functional maximized over $\mathcal{Z}$ uniquely at v. The algorithm is incremental, that is, it computes consecutively the partial zonotopes $\mathcal{Z}^k := \sum_{i=1}^k[-1,1]\cdot z_i$ (or rather, the partial dual hyperplane arrangements). The precise details are quite complicated—indeed, the original proof in [5] was erroneous and was corrected in the later reference [6]. We refer the reader to these two papers for more information and to [7, 11] for some extensions and further applications.

Since our zonotope $\mathcal{P}_w$ is the sum of $m := \binom{n}{2}$ line segments in $\mathbb{R}^d$, these results imply the claimed bound $O(m^{d-1}) = O(n^{2(d-1)})$ on the number of vertices of $\mathcal{P}_w$ and on the computational arithmetic complexity of constructing its vertices and corresponding linear functionals. □

Let $\mathcal{P}^M := \text{conv}\{\mathbf{1}_B \,:\, B\in\mathcal{B}\} \subset \mathbb{R}^n$ be the basis polytope of the matroid $M=(N,\mathcal{B})$, where $\mathbf{1}_B := \sum_{j\in B} e_j$ is the incidence vector of $B\in\mathcal{B}$ with e_j the jth standard unit vector in $\mathbb{R}^n$. We include the short proof of the following statement.

PROPOSITION 2.2. *Every edge of the basis polytope is parallel to e_i-e_j for some pair $i,j\in N$.*

Proof. Consider any pair $A,B\in\mathcal{B}$ of bases such that $[\mathbf{1}_A,\mathbf{1}_B]$ is an edge (that is, a 1-face) of $\mathcal{P}^M$, and let $a\in\mathbb{R}^n$ be a linear functional maximized over $\mathcal{P}^M$ uniquely on that edge. If $A\setminus B=\{i\}$ is a singleton, then $B\setminus A=\{j\}$ is a singleton as well, in which case $\mathbf{1}_A-\mathbf{1}_B=e_i-e_j$ and we are done. Suppose then, indirectly, that it is not, and pick an element i in the symmetric difference $A\Delta B := (A\setminus B)\cup(B\setminus A)$ of A and B of minimum value a_i. Without loss of generality assume $i\in A\setminus B$. Then there is a $j\in B\setminus A$ such that $C := A\setminus\{i\}\cup\{j\}$ is a basis of M. Since $|A\Delta B|>2$, C is neither A nor B. By the choice of i, this basis satisfies $a\cdot\mathbf{1}_C = a\cdot\mathbf{1}_A - a_i + a_j \ge a\cdot\mathbf{1}_A$, and hence $\mathbf{1}_C$ is also a maximizer of a over $\mathcal{P}^M$ and so lies in the 1-face $[\mathbf{1}_A,\mathbf{1}_B]$. But no $\{0,1\}$-vector is a convex combination of others, yielding a contradiction. □

The *normal cone of a face of a polyhedron* $\mathcal{P}$ in $\mathbb{R}^d$ is the relatively open cone of those linear functionals $a\in\mathbb{R}^d$ maximized over $\mathcal{P}$ uniquely on that face. The collection of normal cones of all faces of $\mathcal{P}$ is called the *normal fan* of $\mathcal{P}$. A polyhedron $\mathcal{Z}$ is a *refinement* of a polyhedron $\mathcal{P}$ if the normal fan of $\mathcal{Z}$ is a refinement of that of $\mathcal{P}$, that is, the closure of each normal cone of $\mathcal{P}$ is the union of closures of normal cones of $\mathcal{Z}$. We have the following lemma.

LEMMA 2.3. *The zonotope $\mathcal{P}_w$ is a refinement of the polytope $\mathcal{P}_w^M$.*

Proof. It is known that if $\mathcal{Z} = \sum_{i=1}^{m}[-1,1]\cdot z_i$ and every edge of a polytope $\mathcal{P}$ is parallel to some z_i, then $\mathcal{Z}$ refines $\mathcal{P}$. To see this, consider any vertex u of $\mathcal{Z}$. Then $u = \sum_{i=1}^{m}\lambda_i z_i$ for some $\lambda_i = \pm 1$, and hence its normal cone consists of those a satisfying $a\cdot\lambda_i z_i > 0$ for all i. Let v be a vertex of $\mathcal{P}$ at which some such $\hat{a}$ is maximized. Consider any edge $[v,w]$ of $\mathcal{P}$. Then $v - w = \alpha_i z_i$ for some scalar $\alpha_i \neq 0$ and some z_i, and $0 \leq \hat{a}\cdot(v-w) = \hat{a}\cdot\alpha_i z_i$, implying $\alpha_i\lambda_i > 0$. It follows that every a in the cone of the vertex u of $\mathcal{Z}$ satisfies $a\cdot(v-w) > 0$ for every edge of $\mathcal{P}$ containing v, and therefore a is also in the cone of the vertex v of $\mathcal{P}$. This shows that the normal cone of any vertex of $\mathcal{Z}$ is contained in the normal cone of some vertex of $\mathcal{P}$, and therefore $\mathcal{Z}$ refines $\mathcal{P}$.

Now, let $\pi : \mathbb{R}^n \longrightarrow \mathbb{R}^d : e_j \mapsto w(j)$ be the natural linear projection sending the unit vector e_j corresponding to the matroid element $j \in N$ to the vector $w(j) \in \mathbb{R}^d$. Then for each $B \in \mathcal{B}$ we have $\pi(\mathbf{1}_B) = w(B)$, and hence

$$\mathcal{P}_w^M \;=\; \mathrm{conv}\{\, w(B) \,:\, B \in \mathcal{B}\,\} \;=\; \mathrm{conv}\{\, \pi(\mathbf{1}_B) \,:\, B \in \mathcal{B}\,\} \;=\; \pi(\mathcal{P}^M),$$

so $\mathcal{P}_w^M$ is a projection of $\mathcal{P}^M$. Thus, each edge of $\mathcal{P}_w^M$ is the projection of some edge of $\mathcal{P}^M$ and hence, by Proposition 2.2, is parallel to $\pi(e_i - e_j) = w(i) - w(j)$ for some pair $i, j \in N$. Thus, as explained above, the zonotope $\mathcal{P}_w = \sum_{1\leq i<j\leq n}[-1,1]\cdot(w(i)-w(j))$ refines $\mathcal{P}_w^M$ as claimed. □

We are now in position to prove our theorem.

Proof of Theorem 1.4. Given data M, w, c, the algorithm proceeds with the following steps: first, compute via Proposition 2.1 the list of $O(\,n^{2(d-1)}\,)$ vertices v of $\mathcal{P}_w$, each v along with a linear functional $a(v) \in \mathbb{R}^d$ maximized over $\mathcal{P}_w$ uniquely at v. Second, for each v do the following: let $a := a(v)$ and define the following weighting of matroid elements by scalars:

$$b : M \longrightarrow \mathbb{R} \,:\, j \mapsto a\cdot w(j) = \sum_{i=1}^{d} a_i w(j)_i;$$

now apply a greedy algorithm to obtain a basis $B(v) \in \mathcal{B}$ of maximum weight $b(B)$, that is, sort N by decreasing b-value (using $O(n\log n)$ operations) and find, using at most n calls to the independence oracle presenting M, the lexicographically first basis $B(v)$. Third, for each v compute the value $c(w(B(v)))$ using the evaluation oracle presenting c; an optimal basis for the convex matroid optimization problem is any $B(v)$ achieving maximal such value among the bases $B(v)$ of vertices v of $\mathcal{P}_w$. The complexity is dominated by the second step, which takes $O(n\log n)$ operations and queries and is repeated $O(\,n^{2(d-1)}\,)$ times, giving the claimed bound.

We now justify the algorithm. First, we claim that each vertex u of $\mathcal{P}_w^M$ satisfies $u = w(B(v))$ for some $B(v)$ produced in the second step of the algorithm. Consider any such vertex u. Since $\mathcal{P}_w$ refines $\mathcal{P}_w^M$ by Lemma 2.3, the normal cone of the vertex u of $\mathcal{P}_w^M$ contains the normal cone of some (possibly more than one) vertex v of $\mathcal{P}_w$. Then $a := a(v)$ is maximized over $\mathcal{P}_w^M$ uniquely at u. Now, consider the second step of the algorithm applied to v, and let b be the corresponding scalar weighting of matroid elements. Then the b-weight of any basis B satisfies

$$b(B) \;=\; \sum_{j\in B} a\cdot w(j) \;=\; a\cdot\sum_{j\in B} w(j) \;=\; a\cdot w(B) \;\leq\; a\cdot u$$

with equality if and only if $w(B) = u$. Thus, the maximum b-weight basis $B(v)$ produced by the greedy algorithm will satisfy $u = w(B(v))$. Thus, as claimed, each vertex u of $\mathcal{P}_w^M$ is obtained as $u = w(B(v))$ for some $B(v)$.

Now, since c is convex, the maximum value $c(w(B))$ of any basis $B \in \mathcal{B}$ will occur at some vertex $u = w(B(v))$ of $\mathcal{P}_w^M = \text{conv}\{\, w(B) \,:\, B \in \mathcal{B}\,\}$. Therefore, any basis $B(v)$ with maximum value $c(w(B(v)))$ is an optimal solution to the convex matroid optimization problem. The third step of the algorithm produces such a basis, and so the algorithm is justified. □

The use of the refining zonotope $\mathcal{P}_w$ enhances the efficient enumeration of the vertices of the polytope $\mathcal{P}_w^M = \text{conv}\{\, w(B) \,:\, B \in \mathcal{B}\,\}$ by the algorithm underlying the proof Theorem 1.4. In particular, it shows that the number of vertices of $\mathcal{P}_w^M$, which is a projection of the basis polytope of M, is $O(n^{2(d-1)})$. Since the dimension d is assumed to be fixed, and since each facet of $\mathcal{P}_w^M$ is determined by some d vertices, this implies at once that all facets of $\mathcal{P}_w^M$ can be enumerated in arithmetic complexity $n^{o(d^2)}$. However, as pointed out by one of the referees, the facets can be enumerated directly in complexity $n^{o(d)}$ as well; this procedure can, in turn, be used to enumerate the vertices, albeit in complexity $n^{o(d^2)}$ versus the $n^{o(d)}$ bound guaranteed by Theorem 1.4.

Acknowledgment. Many thanks to the referees for useful suggestions that improved the presentation of this article.

REFERENCES

[1] N. Alon and S. Onn, *Separable partitions*, Discrete Appl. Math., 91 (1999), pp. 39–51.

[2] S. Aviran and S. Onn, *Momentopes and the vertex complexity of partition polytopes*, Discrete Comput. Geom., 27 (2002), pp. 409–417.

[3] K. Allemand, K. Fukuda, T. M. Liebling, and E. Steiner, *A polynomial case of unconstrained zero-one quadratic optimization*, Math. Program., 91 (2001), pp. 49–52.

[4] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver, *Combinatorial Optimization*, John Wiley, New York, 1997.

[5] H. Edelsbrunner, J. O'Rourke, and R. Seidel, *Constructing arrangements of lines and hyperplanes with applications*, SIAM J. Comput., 15 (1986), pp. 341–363.

[6] H. Edelsbrunner, R. Seidel, and M. Sharir, *On the zone theorem for hyperplane arrangements*, in New Results and Trends in Computer Science, Lecture Notes in Comput. Sci. 555, Springer, Berlin, 1991, pp. 108–123.

[7] P. Gritzmann and B. Sturmfels, *Minkowski addition of polytopes: Complexity and applications to Gröbner bases*, SIAM J. Discrete Math., 6 (1993), pp. 246–269.

[8] L. P. Hammer, P. Hansen, P. M. Pardalos, and D. J. Rader, *Maximizing the Product of Two Linear Functions in* $0 - 1$ *Variables*, RUTCOR research report 2-97, Rutgers University, Piscataway, NJ, 1997.

[9] E. F. Harding, *The number of partitions of a set of* n *points in* k *dimensions induced by hyperplanes*, Proc. Edinburgh Math. Soc. (2), 15 (1966/1967), pp. 285–289.

[10] F. K. Hwang, S. Onn, and U. G. Rothblum, *A polynomial time algorithm for shaped partition problems*, SIAM J. Optim., 10 (1999), pp. 70–81.

[11] S. Onn and L. J. Schulman, *The vector partition problem for convex objective functions*, Math. Oper. Res., 26 (2001), pp. 583–590.

SIAM J. DISCRETE MATH.
Vol. 17, No. 2, pp. 254–263

SYMMETRIC CODES OVER RINGS*

ANDRÉ BARBÉ[†] AND FRITZ VON HAESELER[†]

Abstract. Shannon suggested investigating a binary multiplying channel and asked for a maximal uniquely decodable symmetric code. Motivated by his example, we define such a code, called a Shannon set, for an arbitrary ring. We investigate the Shannon sets for the rings $\mathbb{F}_q^{n\times n}$ for $n \in \mathbb{N}$, and the rings $\mathbb{Z}_m = \mathbb{Z}/(m\mathbb{Z})$ for $m \in \mathbb{N}$.

Key words. zero-error information theory, two way communication

AMS subject classification. 94A

DOI. 10.1137/S0895480102413771

1. Introduction. In 1961, Shannon [3] formulated the following problem. Two terminals, $T1$ and $T2$, wish to communicate over a binary multiplicative channel. To this end, they choose a code, i.e., a set $\mathbf{X} \subseteq \{0,1\}^n$, which is also the set of input vectors to the channel. If $\underline{x}, \underline{y} \in \mathbf{X}$ are input vectors, then the output vector is given by $\underline{z} = (z_i)_{i=1,\ldots,n} = (x_i y_i)_{i=1,\ldots,n}$ (the usual product of natural numbers). Each terminal should be able to determine uniquely the vector transmitted by the other terminal on the basis of its own vector and the output vector. Since both terminals are using the same set $\mathbf{X}$, the set is called a symmetric code.

Shannon's question was to determine the possible sets $\mathbf{X}$ and to find a set with highest possible cardinality. A set $\mathbf{X}$ with this property is called uniquely decodable. This problem and its generalizations have been studied by several authors; see, e.g., [1, 4] for more references.

We can reformulate Shannon's question in a ring theoretic setting by considering $\{0,1\}^n$ as the ring $\mathcal{R} = \mathbb{Z}_2^n$, i.e., the n-fold product of the ring $\mathbb{Z}_2 = \mathbb{Z}/(2\mathbb{Z})$. Then Shannon's problem comes down to finding a subset $\mathbf{X} \subset \mathcal{R}$ such that for all given x, $y \in \mathbf{X}$ the knowledge of x and xy allows one to find $y \in \mathbf{X}$ uniquely, and similarly, a knowledge of y and xy allows one to determine $x \in \mathbf{X}$ uniquely. This subset $\mathbf{X}$ is then said to have the unique decoding property.

In this paper we study Shannon's question for arbitrary rings $\mathcal{R}$. We introduce the notion of a Shannon set of a ring $\mathcal{R}$ as a subset with the unique decoding property which has maximal cardinality. The cardinality of a Shannon set is called the Shannon number of the ring.

After some preliminary definitions and some general observations, we will study the ring $\mathbb{F}_q^{n\times n}$, i.e., the ring of $n \times n$-matrices with entries in the finite field $\mathbb{F}_q$ of characteristic p, where $q = p^\alpha$ is the number of elements of $\mathbb{F}_q$ and p is a prime number. We shall prove that the set of invertible matrices is the only Shannon set of the ring $\mathbb{F}_q^{n\times n}$.

*Received by the editors September 5, 2002; accepted for publication (in revised form) May 21, 2003; published electronically November 14, 2003. This paper is an extended version of a paper presented at the 23rd Symposium on Information Theory in the Benelux, Louvain-la-Neuve, Belgium, May 29–31, 2002. This research was supported by the Concerted Action Project GOA-Mefisto of the Flemisch Community, by the FWO (Fund for Scientific Research Flanders) projects G.0080.01, G.0197.02, G.0115.03, and by the Belgian Program on Interuniversity Attraction Poles of the Belgian Prime Minister's Office for Science Technology and Culture (IAP V-22).

http://www.siam.org/journals/sidma/17-2/41377.html

[†]Departement Elektrotechniek (ESAT-SISTA/COSIC), Kasteelpark Arenberg 10, B 3001 Leuven, Belgium (Andre.Barbe@esat.kuleuven.ac.be, fvanhaes@esat.kuleuven.ac.be).

We shall also study the ring $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z}$, i.e., the set of integers modulo m with addition and multiplication modulo m. We shall show that the Shannon number of $\mathbb{Z}/(m)$ is equal to $\phi(m)$, where ϕ is Euler's totient function. However, contrary to the result in the case of the ring $\mathbb{F}_q^{n\times n}$, there may be several Shannon sets. Table 1 provides a complete list.

Table 1
The number of different Shannon sets for the ring $\mathbb{Z}_m$.

Properties of m	Number of Shannon sets
$\gcd(m,6)=1$	1
$m=2m'$ and $\gcd(m',6)=1$	$2^{\phi(m')}$
$m=4m'$ and $\gcd(m',2)=1$	$2^{\phi(m')}+1$
$m=2^{\alpha}m'$, $\alpha\geq 3$, and $\gcd(m',6)=1$	1
$m=3^{\beta}m'$, $\beta\geq 1$, and $\gcd(m',6)=1$	1
$m=6m'$ and $\gcd(m',6)=1$	$2^{2\phi(m')+1}-2^{\phi(m')}$
$m=2\cdot 3^{\beta}m'$, $\beta\geq 2$, and $\gcd(m',6)=1$	$2^{3^{\beta}\phi(m')}$
$m=2^{\alpha}3^{\beta}m'$, $\alpha\geq 3$, $\beta\geq 1$, and $\gcd(m',6)=1$	1

2. Shannon sets of a ring. From now on, $\mathcal{R}$ denotes a ring with 1 and 0. For $r\in\mathcal{R}$ and a subset $A\subset\mathcal{R}$, we write rA as abbreviation for the set $\{ra \mid a\in A\}$. The cardinality of A is denoted by $|A|$. For $s\in\mathcal{R}$ we define $l_s:\mathcal{R}\to\mathcal{R}$ and $r_s:\mathcal{R}\to\mathcal{R}$ as

$$l_s(t)=st \text{ and } r_s(t)=ts,$$

respectively.

Definition 2.1. *A subset* $\mathbf{X}$ *of* $\mathcal{R}$ *has the* unique decoding property *if for every* $s\in\mathbf{X}$ *the maps* $l_s:\mathbf{X}\to\mathcal{R}$ *and* $r_s:\mathbf{X}\to\mathcal{R}$ *are injective, respectively.*

In case of a finite ring $\mathcal{R}$ one has the following obvious lemma.

Lemma 2.2. *Let* $\mathcal{R}$ *be a finite ring. A subset* $\mathbf{X}$ *of* $\mathcal{R}$ *has the* unique decoding property *if and only if* $|r\mathbf{X}|=|\mathbf{X}r|=|\mathbf{X}|$ *holds for all* $r\in\mathbf{X}$.

In the case that $\mathcal{R}$ is commutative, it is sufficient to consider either l_s or r_s. If, furthermore, $\mathcal{R}$ is finite, then a subset $\mathbf{X}$ of $\mathcal{R}$ has the unique decoding property if $|x\mathbf{X}|=|\mathbf{X}|$ for all $x\in\mathbf{X}$.

Note that the problem of finding a subset $\mathbf{X}$ satisfying the requirements of Definition 2.1 is equivalent to Shannon's problem mentioned in the introduction. Indeed, if terminal T1 knows $x\in\mathbf{X}$ and $xy\in\mathcal{R}$, then, due to the fact that $l_x:\mathbf{X}\to\mathcal{R}$ is injective, it can uniquely determine $y\in\mathbf{X}$. The same holds for terminal T2 and its knowledge of $y\in\mathbf{X}$ and $xy\in\mathcal{R}$.

The set of all subsets of $\mathcal{R}$ having the unique decoding property is denoted $\mathcal{U}(\mathcal{R})$. If $x\in\mathcal{R}$, then $\{x\}\in\mathcal{U}(\mathcal{R})$. We state some elementary properties without proof.

Lemma 2.3.

- *If* $\mathbf{X}\in\mathcal{U}(\mathcal{R})$, *then any subset of* $\mathbf{X}$ *has the unique decoding property.*
- *If* $\mathbf{X}_1$, $\mathbf{X}_2\in\mathcal{U}(\mathcal{R})$, *then* $\mathbf{X}_1\cap\mathbf{X}_2\in\mathcal{U}(\mathcal{R})$.
- *If* $\mathcal{R}_1$ *and* $\mathcal{R}_2$ *are rings, then* $\mathcal{U}(\mathcal{R}_1)\times\mathcal{U}(\mathcal{R}_2)\subset\mathcal{U}(\mathcal{R}_1\times\mathcal{R}_2)$.

The relation $\subset$ defines a partial order on $\mathcal{U}(\mathcal{R})$. It is therefore meaningful to speak of maximal elements of $\mathcal{U}(\mathcal{R})$. In the case of a finite ring certain maximal elements in $\mathcal{U}(\mathcal{R})$ are special.

Definition 2.4. *Let* $\mathcal{R}$ *be a finite ring. The set* $\mathbf{X}\in\mathcal{U}(\mathcal{R})$ *is called a* Shannon set *of the finite ring* $\mathcal{R}$ *if*

$$|\mathbf{X}|=\max\{|Y| \mid Y\in\mathcal{U}(\mathcal{R})\}.$$

The cardinality $|\mathbf{X}|$ *of a Shannon set is called the* Shannon number *of the ring* $\mathcal{R}$ *and is denoted as* $\chi(\mathcal{R})$.

In particular, Shannon sets of $\mathcal{R}$ are maximal in $\mathcal{U}(\mathcal{R})$. On the other hand, of course, not every maximal element of $\mathcal{U}(\mathcal{R})$ is a Shannon set. For example, the set $\{0\}$ is a set with the unique decomposition property and $\{0\}$ is maximal in $\mathcal{U}(\mathcal{R})$.

The following lemma provides a lower bound for the Shannon number of a ring.

LEMMA 2.5. *If* $\mathcal{R}^\star$ *denotes the set of units of the finite ring* $\mathcal{R}$, *then* $\chi(\mathcal{R}) \geq |\mathcal{R}^*|$.

The proof is obvious. As the ring $(\mathbb{Z}_2)^n$ shows, this estimate can be very poor. In fact, we have only one unit and large sets with the unique decoding property (see, e.g., [4]). If a finite commutative ring with 0 and 1 has large Shannon sets, then it is already a field, as shown by the following theorem.

THEOREM 2.6. *Let* $\mathcal{R}$ *be a finite commutative ring with* 0 *and* 1. $\mathcal{R}$ *is a field if and only if* $\chi(\mathcal{R}) = |\mathcal{R}| - 1$.

Proof. If $\mathcal{R}$ is a field, then the assertion is obviously true.

If $\chi(\mathcal{R}) = |\mathcal{R}| - 1$, then for any $r \neq 0$ the map $l_r : \mathcal{R} \setminus \{0\} \to \mathcal{R}$, $x \mapsto rx$ is injective. Now suppose that there exist r, $s \in \mathcal{R} \setminus \{0\}$ such that $rs = 0$. We compute $r(s+1) = rs + r = r \cdot 1$ and conclude that, due to the injectivity property of l_r, it follows that either $s + 1 = 0$ or $s + 1 = 1$. If $s + 1 = 0$, then it follows that $r = 0$, and if $s + 1 = 1$, then it follows that $s = 0$; this contradicts our assumption. Therefore any map l_r restricted to $\mathcal{R} \setminus \{0\}$ is injective with image $\mathcal{R} \setminus \{0\}$; thus $\mathcal{R}$ is a field. □

We conclude the section with some examples.

1. For $m \geq 2$ we define a ring structure on $\{0, \dots, m-1\}$ by using addition mod m and multiplication given by $rs = 0$ for all r, s. The thus defined ring has Shannon number 1. Any $r \neq 0$ defines a Shannon set $\{r\}$.
2. For $m \geq 2$ we define a ring structure on $\{0, \dots, m-1\}$ by using addition mod m and multiplication given by $0r = r0 = 0$ for all r and $r1 = 1r = r$ and $rs = 0$ for all other possibilities. The thus defined ring has Shannon number 2. In fact, if $r \notin \{0, 1\}$, then $\{1, r\}$ is a Shannon set. If r, s are different and both are different from 1 and 0, then $\{r, s\}$ is not a Shannon set.
3. If $\mathcal{R} = \mathbb{Z}_2^3$, then a Shannon set is given by $\{(0,1,1), (1,0,1), (1,1,0)\}$.
4. Let $\mathcal{R} = \mathbb{Z}$, the ring of integers; then the set $\mathbb{Z}^* = \mathbb{Z} \setminus \{0\}$ has the unique decoding property. Moreover, $\mathbb{Z}^*$ is maximal in the sense that every set $\mathbf{X} \in \mathcal{U}(\mathbb{Z})$, $X \neq \{0\}$, is a subset of $\mathbb{Z}^*$. This observation generalizes to every integral domain, i.e., a commutative ring $\mathcal{R}$ without zero divisors.
5. Let $\mathcal{R} = \mathbb{Z}^{n \times n}$, $n \geq 1$, be the ring of $n \times n$-matrices with entries in $\mathbb{Z}$. Then
$$\mathbf{X} = \left\{ A \mid \det A \neq 0,\ A \in \mathbb{Z}^{n \times n} \right\}$$
belongs to $\mathcal{U}(\mathbb{Z}^{n \times n})$. Furthermore, $\mathbf{X}$ is maximal in $\mathcal{U}(\mathbb{Z}^{n \times n})$.

3. Shannon sets of the ring $\mathbb{F}_q^{n \times n}$. In this section, we study the ring of matrices over a finite field of characteristic p, where p is a prime number. We will show that for rings of this type there exists only one Shannon set.

THEOREM 3.1. *Let* $\mathcal{R} = \mathbb{F}_q^{n \times n}$ *be the ring of* $n \times n$*-matrices with entries in the field* F_q *of characteristic* p. *The set* $Gl(n, \mathbb{F})$ *of invertible matrices is the only Shannon set of* $\mathbb{F}_q^{n \times n}$.

Proof. The set $Gl(n, \mathbb{F}_q)$ of invertible $n \times n$-matrices has the unique decoding property. The cardinality of $Gl(n, \mathbb{F}_q)$ is given by $\prod_{j=0}^{n-1}(q^n - q^j)$. Now suppose that $\mathbf{X} \in \mathcal{U}(\mathbb{F}_q^{n \times n})$ is such that $|\mathbf{X}| > |Gl(n, \mathbb{F}_q)|$. Then $\mathbf{X}$ contains at least one matrix

A such that $A \notin Gl(n, \mathbb{F}_q)$. If we consider A as a linear map from the $\mathbb{F}_q$-vector space $\mathbb{F}_q^n$ to itself, then $A\left(\mathbb{F}_q^n\right)$ is an at most $(n-1)$-dimensional subspace of $\mathbb{F}_q^n$. This gives that the cardinality of the set $A\mathbb{F}_q^{n\times n} = \{AB \mid B \in \mathbb{F}_q^{n\times n}\}$ is less than or equal to $q^{n(n-1)}$. On the other hand, $A \in \mathbf{X}$, and therefore $|A\mathbf{X}| = |\mathbf{X}|$. This yields $|Gl(n, \mathbb{F}_q)| = \prod_{j=0}^{n-1}(q^n - q^j) < |\mathbf{X}| = |A\mathbf{X}| \leq q^{n(n-1)}$, which is a contradiction.

Therefore, no element of $\mathcal{U}(\mathbb{F}_q^{n\times n})$ has a cardinality greater than $Gl(n, \mathbb{F}_q)$. As the above arguments show, a Shannon set does not contain a noninvertible matrix. This shows that $Gl(n, \mathbb{F}_q)$ is the only Shannon set. □

As a consequence we note the following.

Corollary 3.2.

$$\chi\left(\mathbb{F}_q^{n\times n}\right) = \prod_{j=0}^{n-1}(q^n - q^j).$$

4. Shannon sets of the ring $\mathbb{Z}_m$. In this section we study the residue class rings $\mathbb{Z}_m = \mathbb{Z}/(m\mathbb{Z})$, where m is a natural number. Theorem 4.2 shows that the Shannon number of the ring $\mathbb{Z}_m$ is equal to the cardinality of the set of units $\mathbb{Z}_m^*$. Moreover, we provide the proofs for the results listed in Table 1.

We begin with an auxiliary lemma and some notation.

Lemma 4.1. *If $1 < p_2 < p_3 < \cdots < p_L$ are natural numbers, then*

$$\prod_{j=2}^{L} \frac{p_j}{p_j - 1} \leq p_L,$$

where equality holds if and only if $j = p_j$ for all $j = 2, \ldots, L$.

The simple proof is omitted.

The function $\phi : \mathbb{N} \to \mathbb{N}$ defined as $\phi(m) = m\prod_{p|m}(1 - \frac{1}{p})$, where the product runs over all prime numbers p that divide m, is called *Euler's totient function* (see [2]). $\phi(m)$ is equal to the number of units in the ring $\mathbb{Z}_m$. Moreover, ϕ has the following two properties:

- For all prime numbers and all $\alpha \in \mathbb{N} \setminus \{0\}$,

$$\phi(p^\alpha) = p^{\alpha-1}\phi(p).$$

- If m_1 and m_2 are relatively prime, i.e., $\gcd(m_1, m_2) = 1$, then

$$\phi(m_1 m_2) = \phi(m_1)\phi(m_2).$$

If d is a divisor of m, then the set $Z_m(d)$ denotes the set $\{l \in \mathbb{Z}_m \mid \gcd(l, m) = d\}$, and we have

$$|Z_m(d)| = \phi\left(\frac{m}{d}\right).$$

Moreover, for two divisors d_1 and d_2 of m we have that $d_1 Z_m(d_2) = Z_m(\gcd(d_1 d_2, m))$. The collection of sets $Z_m(d)$, d divides m, forms a partition of $\mathbb{Z}_m$, i.e., $\mathbb{Z}_m = \cup_{d|m} Z_m(d)$, and every $x \in \mathbb{Z}_m$ is contained in a unique $Z_m(d)$.

Moreover, as a consequence of Lemma 4.1, we have the inequality

$$\frac{m}{\phi(m)} = \prod_{p|m} \frac{p}{p-1} \leq p_L, \tag{1}$$

where p_L is the largest prime divisor of m.

Theorem 4.2. *If $\mathcal{R} = \mathbb{Z}_m$, then the set of units $\mathbb{Z}_m^*$ of $\mathbb{Z}_m$ is a Shannon set.*

Proof. We suppose that $\mathbf{X}$ is a Shannon set such that $|\mathbf{X}| > \phi(m) = |\mathbb{Z}_m^*|$ and show that this leads to a contradiction. We begin with two auxiliary results.

1. If $d_1 = 1 < d_2 < \cdots < d_M$ are the divisors of m such that $d_j < \frac{m}{\phi(m)}$ for all $j = 1, \dots, M$, then

$$\mathbf{X} \subseteq \bigcup_{j=1}^{M} Z_m(d_j).$$

Proof. Suppose $x \in \mathbf{X}$ and $x \in Z_m(d)$, where $d \geq \frac{m}{\phi(m)}$; then, due to our assumption on $|\mathbf{X}| > \phi(m)$ and the fact that $\mathbf{X}$ is a Shannon set, we have

$$\phi(m) < |\mathbf{X}| = |x\mathbf{X}| \leq |d\mathbb{Z}_m| = \frac{m}{d} \leq \phi(m),$$

which is a contradiction.

2. If d is a proper divisor of m, i.e., $d > 1$ such that $d < \frac{m}{\phi(m)}$, then $\mathbf{X} \cap Z_m(d) = \emptyset$.

Proof. Let d be the largest proper divisor of m such that $Z_m(d) \cap \mathbf{X} \neq \emptyset$ and let $d_1 = 1 < d_2 < \cdots < d_{K-1} < d_K = d$ be all divisors of m less than or equal to d_K. By 1, it follows that $d_K < \frac{m}{\phi(m)}$. If $m^* = \operatorname{lcm}(d_1, \dots, d_K)$ denotes the least common multiple of $d_1, \dots, d_K$, then we can write

$$m = m^* \prod_{i=1}^{L} p_i^{\beta_i} m',$$

where $p_1 < \cdots < p_L$ are the prime divisors of m^* and the prime divisors of m' are all greater than d_K.

By 1 we have that

$$\mathbf{X} \subseteq Z_m(d_1) \cup \cdots \cup Z_m(d_K).$$

Let $y \in \mathbf{X} \cap Z_m(d)$. Since $\mathbf{X}$ is a Shannon set of cardinality greater than $\phi(m)$, it follows that

$$\phi(m) < |\mathbf{X}| = |y\mathbf{X}| \leq |d_K \left(Z_m(d_1) \cup \cdots \cup Z_m(d_K) \right)|,$$

which implies

$$\phi(m) < \left| \bigcup_{j=1}^{K} Z_m(\gcd(d_K d_j, m)) \right| = \left| \bigcup_{j=1}^{K} Z_m \left(d_K \gcd \left(d_j, \frac{m^*}{d_K} \prod_{i=1}^{L} p_i^{\beta_i} m' \right) \right) \right|.$$

This can be estimated by

$$\phi(m) < \left| \bigcup_{j=1}^{K} Z_m \left(d_K \gcd \left(d_j, \frac{m^*}{d_K} \prod_{i=1}^{L} p_i^{\beta_i} m' \right) \right) \right| \leq \left| \bigcup_{\delta} Z_m(d_K \delta) \right|,$$

where δ runs over the divisors of $\frac{m^*}{d_K} \prod_{i=1}^{L} p_i^{\beta_i}$. Using Euler's totient function ϕ, we obtain the inequality

$$\phi(m) < \left| \bigcup_{\delta} Z_m(d_K \delta) \right| = \sum_{\delta} \phi \left(\frac{m}{d_K \delta} \right) = \phi(m') \sum_{\delta} \phi \left(\frac{m^* \prod_{i=1}^{L} p_i^{\beta_i}}{d_K \delta} \right).$$

Since the sum is over all divisors of $\frac{m^*}{d_K}\prod_{i=1}^{L}p_i^{\beta_i}$, we obtain the inequality

$$\phi(m) < \phi(m')\frac{m^*}{d_K}\prod_{i=1}^{L}p_i^{\beta_i}.$$

With $\phi(m) = \phi(m^*\prod_{i=1}^{L}p_i^{\beta_i}m') = \phi(m^*\prod_{i=1}^{L}p_i^{\beta_i})\phi(m')$ and the fact that

$$\phi\left(m^*\prod_{i=1}^{L}p_i^{\beta_i}\right) = m^*\prod_{i=1}^{L}p_i^{\beta_i}\prod_{i=1}^{L}\left(1-\frac{1}{p_i}\right) = \phi(m^*)\prod_{i=1}^{L}p_i^{\beta_i},$$

we obtain the final inequality

$$\phi(m^*) < \frac{m^*}{d_K}$$

or $d_K < \frac{m^*}{\phi(m^*)}$. Since $p_1 < \cdots < p_L$ are the prime divisors of m^*, there exists a $j_0 \in \{1, \ldots, K\}$ such that $d_{j_0} = p_L \leq d_K$, and thus we have (see (1)),

$$p_L \leq d_K < \frac{m^*}{\phi(m^*)} = \prod_{j=1}^{L}\frac{p_j}{p_j-1},$$

which contradicts Lemma 4.1. Therefore we can conclude that no proper divisor d of m exists such that $\mathbf{X} \cap Z_m(d) \neq \emptyset$. This proves the second auxiliary result.

We are now prepared to finish the proof of Theorem 4.2. By 1, we have $\mathbf{X} \subseteq Z_m(d_1) \cup \cdots \cup Z_m(d_M)$, where d_M is the largest divisor less than $m/\phi(m)$. By 2, it follows that $\mathbf{X} \subseteq Z_m(1)$, which contradicts our assumption that $|\mathbf{X}| > \phi(m)$. This completes the proof of Theorem 4.2. □

THEOREM 4.3. *If* $\gcd(m, 6) = 1$, *then* $\mathbb{Z}_m^*$ *is the unique Shannon set of the ring* $\mathbb{Z}_m$.

Proof. If $\mathbf{X}$ is a Shannon set different from $\mathbb{Z}_m^*$, then there exists a largest nontrivial divisor d of m such that $\mathbf{X} \cap Z_m(d) \neq \emptyset$. Due to the above observations in 1 of the previous proof, d has to satisfy $d \leq \frac{m}{\phi(m)}$. Following the arguments given in 2 of the proof of Theorem 4.2, d also satisfies

$$d \leq \frac{m^*}{\phi(m^*)},$$

where m^* is the lowest common multiple of all divisors of m which are less than or equal to d. If $p_2 < \cdots < p_L$ denote the prime divisors of m^*, then we have that

$$p_L \leq d \leq \frac{m^*}{\phi(m^*)} = \prod_{j=2}^{L}\frac{p_j}{p_j-1}. \tag{2}$$

Since $\gcd(m, 6) = 1$, we have $4 < p_2 < \cdots < p_L$ and Lemma 4.1 implies

$$\prod_{j=2}^{L}\frac{p_j}{p_j-1} < p_L,$$

which contradicts inequality (2). □

The proof of Theorem 4.3 gives a hint of how to construct Shannon sets $\mathbf{X} \subset \mathbb{Z}_m$ such that $\mathbf{X} \neq Z_m(1)$. If $\mathbf{X}$ is a Shannon set and d a maximal divisor of m such that $\mathbf{X} \cap Z_m(d) \neq \emptyset$, then our above considerations yield the inequalities

$$\frac{m^*}{\phi(m^*)} \le p_L \le d \le \frac{m^*}{\phi(m^*)},$$

where $m^* = \operatorname{lcm}(d_1, \ldots, d_K = d)$ and $p_1 < \cdots < p_L$ are the prime divisors of m^*. It follows that $p_L = m^*/\phi(m^*)$ and, due to inequality (1), it follows that the prime divisors of m^* may be either $p_1 = 2$ or $p_1 = 2 < p_2 = 3$. This implies that $\mathbf{X} \cap Z_m(d) \neq \emptyset$ is only possible for $d = 2, 3$.

For natural numbers m such that $\gcd(m, 6) \neq 1$, the existence of a unique Shannon set depends on divisibility properties of m w.r.t. 2 and 3.

THEOREM 4.4. *Let $m = 2^\alpha 3^\beta m'$ such that α, $\beta \in \mathbb{N}$ and $\gcd(m', 6) = 1$.*

1. *If $\alpha = 0$ and $\beta \ge 1$, then $Z_m(1)$ is the only Shannon set of $\mathbb{Z}_m$.*
2. *If $\alpha \ge 3$ and $\beta \ge 0$, then $Z_m(1)$ is the only Shannon set of $\mathbb{Z}_m$.*

Proof. 1. A Shannon set $\mathbf{X}$ different from $Z_m(1)$ satisfies $\mathbf{X} \cap Z_m(3) \neq \emptyset$. In particular, $\mathbf{X} \subset Z_m(1) \cup Z_m(3)$.

For $\beta = 1$ and $x \in \mathbf{X} \cap Z_m(3)$, we have the inequality

$$\phi(m) = \phi(3m') = |x\mathbf{X}| \le |x(Z_m(1) \cup Z_m(3))| = |Z_m(3)|.$$

By the elementary properties of ϕ, one obtains $\phi(3m') = 2\phi(m') \le \phi(m')$, which is a contradiction.

For $\beta \ge 2$, i.e., $m = 3^\beta m'$, the assumption $x \in \mathbf{X} \cap Z_m(3)$ leads to the inequality

$$\phi(m) = |x\mathbf{X}| \le |x(Z_m(1) \cup Z_m(3))| \le |Z_m(3) \cup Z_m(9)|,$$

which gives $\phi(3^\beta m') \le \phi(3^{\beta-1} m') + \phi(3^{\beta-2} m')$. This yields

$$\phi(3^\beta) \le \phi(3^{\beta-1}) + \phi(3^{\beta-2}),$$

which is equivalent to

$$3^2 \phi(3^{\beta-2}) \le 3\phi(3^{\beta-2}) + \phi(3^{\beta-2}).$$

This leads to the contradiction $9 \le 3 + 1$.

2. We start with $\beta = 0$, i.e., $m = 2^\alpha m'$, $\alpha \ge 3$. If $\mathbf{X}$ is a Shannon set such that $x \in \mathbf{X} \cap Z_m(2)$, then, by similar arguments as in 1, we obtain the inequality

$$\phi(2^\alpha) \le \phi(2^{\alpha-1}) + \phi(2^{\alpha-2})$$

which gives $4 \le 3$, which is a contradiction.

For $\beta = 1$, i.e., $m = 2^\alpha 3m'$, a Shannon set $\mathbf{X}$ different from $Z_m(1)$ satisfies $\mathbf{X} \cap (Z_m(2) \cup Z_m(3)) \neq \emptyset$. Suppose there exists a $x \in \mathbf{X} \cap Z_m(3)$; then we have $\phi(m) = |\mathbf{X}| = |x\mathbf{X}| \le |Z_m(3) \cup Z_m(6)|$, which gives the inequality

$$\phi(2^\alpha 3) \le \phi(2^\alpha) + \phi(2^{\alpha-1}),$$

which yields $4 \le 3$, which is a contradiction. Thus we conclude that $\mathbf{X}$ satisfies $\mathbf{X} \cap Z_m(3) = \emptyset$.

For $x \in \mathbf{X} \cap Z_m(2)$ we obtain $x\mathbf{X} \subset Z_m(2) \cup Z_m(4)$, which yields that the estimate

$$\phi(2^\alpha 3) \le \phi(2^{\alpha-1} 3) + \phi(2^{\alpha-2} 3)$$

is not true for all $\alpha \ge 3$.

Finally, we consider $\beta \geq 2$, i.e., $m = 2^\alpha\, 3^\beta m'$. Then the assumption $x \in \mathbf{X} \cap Z_m(3)$ leads to the inequality

$$\phi(2^\alpha\, 3^\beta) \leq \phi(2^\alpha\, 3^{\beta-1}) + \phi(2^{\alpha-1}\, 3^{\beta-1}) + \phi(2^\alpha\, 3^{\beta-2}),$$

which is equivalent to $6 \leq 3 + \frac{\phi(3^{\beta-2})}{3^{\beta-2}} \leq 4$, which is a contradiction. The assumption $x \in \mathbf{X} \cap Z_m(2)$ leads again to a contradiction. □

In the remaining cases, i.e., $m = 2m'$, $\gcd(m', 2) = 1$ and $m = 4m'$, $\gcd(m', 2) = 1$, there exists more than one Shannon set.

THEOREM 4.5. *For a natural number of the form $m = 2^\alpha m'$, where $\alpha = 1, 2$ and $\gcd(m', 2) = 1$, the following holds.*

1. *If $\alpha = 1$ and $m = 2 \cdot 3^\beta m'$, where $\gcd(6, m') = 1$ and $\beta \neq 1$, then there exist $2^{\phi(3^\beta m')}$ different Shannon sets of $\mathbb{Z}_m$.*
2. *If $\alpha = 1$ and $m = 6m'$, where $\gcd(m', 6) = 1$, then there exist $2^{2\phi(m')+1} - 2^{\phi(m')}$ different Shannon sets of $\mathbb{Z}_m$.*
3. *If $\alpha = 2$, then there exist $2^{\phi(m')} + 1$ different Shannon sets of $\mathbb{Z}_m$.*

Proof. 1 (a) We begin with the case $\beta = 0$, i.e., $m = 2m'$. By the arguments given in the proof of Theorem 4.3, a Shannon set $\mathbf{X}$ has to be a subset of $Z_m(1) \cup Z_m(2)$, and we write

$$\mathbf{X} = \mathbf{X}(1) \cup \mathbf{X}(2),$$

where $\mathbf{X}(i) = \mathbf{X} \cap Z_m(i)$ for $i = 1, 2$. Note that $|Z_m(2)| = |Z_m(1)| = \phi(m)$. Moreover, since $2Z_m(2) = Z_m(2)$, it follows that $Z_m(2)$ has the unique decoding property. Since the cardinality of $Z_m(2)$ equals the cardinality of $Z_m(1)$, it follows that $Z_m(2)$ is a Shannon set. Finally, we note that the map $Z_m(2) \ni x \mapsto x + m' \in Z_m(1)$ is bijective.

Let $\mathbf{X}$ be a Shannon set and let $x \in \mathbf{X}(2)$. Since $\mathbf{X}$ is a Shannon set, it follows that $x^2 \in x\mathbf{X}$. On the other hand, one has that $x(x + m') = x^2$, and $x + m'$, being an element of $Z_m(1)$, is therefore not an element of $\mathbf{X}(1)$. Thus we see that $\mathbf{X}(1) = Z_m(1) \setminus (\mathbf{X}(2) + m')$ and $\mathbf{X} = (Z_m(1) \setminus (\mathbf{X}(2) + m')) \cup \mathbf{X}(2)$.

On the other hand, if $\mathbf{Y}$ is any subset of $Z_m(2)$, then $(Z_m(1) \setminus (\mathbf{Y} + m') \cup \mathbf{Y}$ is a Shannon set. Therefore, there are $2^{|Z_m(2)|}$ Shannon sets for $\beta = 0$.

(b) Now suppose that $m = 2 \cdot 3^\beta m'$, where $\beta \geq 2$. A Shannon set $\mathbf{X}$ is a subset of $Z_m(1) \cup Z_m(2) \cup Z_m(3)$, and we write

$$\mathbf{X} = \mathbf{X}(1) \cup \mathbf{X}(2) \cup \mathbf{X}(3),$$

where $\mathbf{X}(i) = \mathbf{X} \cap Z(i)$, $i = 1, 2, 3$.

If $x \in \mathbf{X}(3)$, then it follows that $\phi(m) = |x\mathbf{X}| \leq |Z_m(3) \cup Z_m(6) \cup Z_m(9)|$. We therefore obtain the inequality

$$\phi(2 \cdot 3^\beta m') \leq \phi(2 \cdot 3^{\beta-1} m') + \phi(3^{\beta-1} m') + \phi(2 \cdot 3^{\beta-2} m'),$$

which is impossible. This shows that $\mathbf{X} \subset Z_m(1) \cup Z_m(2)$ and an application of the same arguments as above gives the desired result.

2. For $m = 6m'$, where $\gcd(m', 6) = 1$, we have the following facts.

- $Z_m(1) = Z_m(2) + 3m'$.
- $Z_m(1) = (Z_m(3) + 2m') \cup (Z_m(3) + 4m')$ and $(Z_m(3) + 2m') \cap (Z_m(3) + 4m') = \emptyset$.
- $|Z_m(1)| = |Z_m(2)| = 2\phi(m')$ and $|Z_m(3)| = \phi(m')$.

As before, we have $\mathbf{X} = \mathbf{X}(1) \cup \mathbf{X}(2) \cup \mathbf{X}(3)$. If $\mathbf{X}(3) = \emptyset$, then the arguments of 1 (a) for $\beta = 0$ apply and, since $m = 2(3m')$, we count $2^{\phi(3m')} = 2^{2\phi(m')}$ Shannon sets.

If $x \in \mathbf{X}(3)$, then $\mathbf{X}(1)$ does not contain $x + 2m'$ and $x + 4m'$, since $x^2 = x(x + 2m') = x(x + 4m')$. Moreover, if $x' \in \mathbf{X}(1)$, then there exists $\zeta \in Z_m(3)$ such that either $x' = \zeta + 2m'$ or $x' = \zeta + 4m'$. On the other hand, for $x \in \mathbf{X}(3)$ we have that $xx' = x(\zeta + 2m') = x(\zeta + 4m')$. That is, $x' = \zeta + 2m' \in \mathbf{X}(1)$ if and only if $\zeta + 4m' \notin \mathbf{X}(1)$. We conclude that $|\mathbf{X}(1)| \leq \phi(m') - |\mathbf{X}(3)|$. On the other hand, we have that

$$|\mathbf{X}| = 2\phi(m') = |\mathbf{X}(1)| + |\mathbf{X}(2)| + |\mathbf{X}(3)| \leq \phi(m') + |\mathbf{X}(2)|\,.$$

Therefore $|\mathbf{X}(2)| \geq \phi(m')$.

If $x' \in Z_m(2)$, then either $x' + 2m' \in Z_m(2)$ or $x' + 4m' \in Z_m(2)$. In any case it follows for $x \in \mathbf{X}(3)$ and $x' \in \mathbf{X}(2)$ that either $xx' = x(x' + 2m')$ or $xx' = x(x' + 4m')$. Therefore $|\mathbf{X}(2)| \leq \phi(m')$. By our above inequality, it follows that $|\mathbf{X}(2)| = \phi(m')$.

Finally, if $x' \in \mathbf{X}(2)$, then $x' + 3m' \notin \mathbf{X}(1)$; otherwise $x'x' = x'(x' + 3m')$. This gives that

$$\mathbf{X}(1) = Z_m(1) \setminus ((\mathbf{X}(2) + 3m') \cup (\mathbf{X}(3) + 2m') \cup (\mathbf{X}(3) + 4m'))\,.$$

On the other hand, for a given nonempty subset $\mathbf{Y}(3)$ of $Z(3)$ and a subset $\mathbf{Y}(2) \subset Z(2)$ such that $|\mathbf{Y}(2)| = \phi(m')$ and such that for given $x_1, x_2 \in \mathbf{X}(2)$ with $x_1 \not\equiv x_2 \bmod 2m'$, we define $\mathbf{Y}(1)$ as in the above formula (replace $\mathbf{X}$ by $\mathbf{Y}$). Then $\mathbf{Y}(1) \cup \mathbf{Y}(2) \cup \mathbf{Y}(3)$ is a Shannon set.

This gives $2^{\phi(m')}(2^{\phi(m')} - 1)$ Shannon sets if $\mathbf{X}(3) \neq \emptyset$, plus the number of Shannon sets in case $\mathbf{X}(3) = \emptyset$, which gives a total of $2^{2\phi(m')+1} - 2^{\phi(m')}$ Shannon sets.

3. Let $m = 4 \cdot 3^\beta m'$, where $\gcd(m', 6) = 1$. As a first step we consider the case $\beta = 0$, i.e., $m = 4m'$. We have that $|\mathbf{X}(\mathbf{1})| = 2\,|\mathbf{X}(\mathbf{2})| = 2\,|\mathbf{X}(\mathbf{4})|$. Moreover, since the map $Z_m(2) \ni x \mapsto 2x \in Z_m(4)$ is bijective, it follows that $Z_m(2)$ has the unique decoding property. The map $Z_m(1) \ni x \mapsto x + 2m' \in Z_m(1)$ is bijective.

If $\mathbf{X} = \mathbf{X}(1) \cup \mathbf{X}(2)$ is a Shannon set such that $\mathbf{X}(2) \neq \emptyset$, then $x \in \mathbf{X}(1)$ implies $x + 2m' \notin \mathbf{X}(1)$; otherwise we would have, for $y \in \mathbf{X}(2)$, $yx = y(x + 2m')$. This shows that $|\mathbf{X}(1)| \leq \phi(m')$. Since $\mathbf{X}$ is a Shannon set, it follows that $|\mathbf{X}(2)| = \phi(m')$, i.e., $\mathbf{X}(2) = Z_m(2)$, and $|\mathbf{X}(1)| = \phi(m')$. Note that $\mathbf{X}(1)$ is such that $(\mathbf{X}(1) + 2m') \cap \mathbf{X}(1) = \emptyset$.

Let $\mathbf{Y} \subset Z_m(1)$ be such that $(\mathbf{Y} + 2m') \cap \mathbf{Y} = \emptyset$; then $\mathbf{Y} \cup Z_m(2)$ is a Shannon set. This gives $2^{\phi(m')}$ Shannon sets. Since $Z_m(1)$ is also a Shannon set, the assertion follows.

As a next step we consider $m = 12m'$, where $\gcd(m', 6) = 1$. A Shannon set $\mathbf{X}$ is a subset of $Z_m(1) \cup Z_m(2) \cup Z_m(3)$. If we assume that there exists $x \in \mathbf{X}(3)$, then we arrive at the inequality $\phi(12) \leq \phi(4) + \phi(2)$, which is a contradiction. Thus we conclude that $\mathbf{X} \subset Z_m(1) \cup Z_m(2)$, and we apply the same arguments as for the case $m = 4m'$.

The case $m = 4 \cdot 3^\beta$ for $\beta \geq 2$ is treated in a similar way. The assumption $\mathbf{X}(3) \neq \emptyset$ leads to a contradiction. Thus $\mathbf{X} \subset Z_m(1) \cup Z_m(2)$, and we can use the same arguments as in 1 for $\beta = 0$. □

REFERENCES

[1] R. Ahlswede, N. Cai, and Z. Zhang, *On interactive communication*, IEEE Trans. Inform. Theory, 43 (1997), pp. 22–37.

[2] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 5th ed., Oxford Clarendon Press, Oxford, UK, 1979.

[3] C. E. Shannon, *Two-way communication channels*, in Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics, and Probability, Berkeley, CA, 1961, pp. 611–644.

[4] L. Tolhuizen, *New rate pairs in the zero-error capacity region of the binary multiplying channel without feedback*, IEEE Trans. Inform. Theory, 46 (2000), pp. 1043–1046.

SIAM J. Discrete Math.
Vol. 17, No. 2, pp. 264–275

LABELING PLANAR GRAPHS WITH CONDITIONS ON GIRTH AND DISTANCE TWO*

WEI-FAN WANG[†] AND KO-WEI LIH[‡]

Abstract. For a planar graph G, let $\Delta(G)$, $g(G)$, and $\lambda(G; p, q)$ denote, respectively, its maximum degree, girth, and $L(p, q)$-labeling number. We prove that (1) $\lambda(G; p, q) \leq (2q - 1)\Delta(G) + 4p + 4q - 4$ if $g(G) \geq 7$; (2) $\lambda(G; p, q) \leq (2q - 1)\Delta(G) + 6p + 12q - 9$ if $g(G) \geq 6$; (3) $\lambda(G; p, q) \leq (2q - 1)\Delta(G) + 6p + 24q - 15$ if $g(G) \geq 5$. These bounds have consequences on conjectures by Wegner [*Graphs with Given Diameter and a Coloring Problem*, preprint, University of Dortmund, Dortmund, Germany, 1977] and Griggs and Yeh [*SIAM J. Discrete Math.*, 5 (1992), pp. 586–595].

Key words. $L(p, q)$-labeling, planar graph, girth

AMS subject classification. 05C15

DOI. 10.1137/S0895480101390448

1. Introduction.

All graphs considered in this paper are finite simple graphs. A *plane* graph is a particular drawing of a planar graph on the Euclidean plane. For a plane graph G, let $V(G)$, $E(G)$, $F(G)$, $\Delta(G)$, and $\delta(G)$ denote, respectively, its vertex set, edge set, face set, maximum degree, and minimum degree. The *girth* $g(G)$ of a graph G is the length of a shortest cycle of G.

A coloring of a graph G is a mapping ϕ from $V(G)$ to the set $\{0, 1, \ldots, k-1\}$ for some positive integer k. A coloring is said to be proper if $\phi(x) \neq \phi(y)$ for every edge xy of G. The *chromatic number* $\chi(G)$ is the smallest k such that G has a proper coloring into the set $\{0, 1, \ldots, k-1\}$. The *distance* between two vertices is the length of a shortest path connecting them. The *square* G^2 of a graph G is the graph defined on the vertex set $V(G)$ such that distinct vertices are adjacent in G^2 if and only if their distance is at most 2 in G.

Obviously, $\chi(G^2) \geq \Delta(G) + 1$. A tree of order at least 2 attains this lower bound. Moreover, it is easy to see that $\chi(G^2) \leq \Delta^2(G) + 1$ for any graph G. There exist infinitely many graphs that attain this upper bound. Two of the simplest examples are a cycle of length 5 and the Petersen graph.

Wegner [13] first investigated the chromatic number of the square of a planar graph. He proved that $\chi(G^2) \leq 8$ for every planar graph G with $\Delta(G) = 3$ and conjectured that the upper bound could be reduced to 7. Recently, Thomassen [12] has established Wegner's conjecture. In [13], Wegner also proposed the following conjecture. The upper bounds are sharp if the conjecture is true.

Conjecture 1. *Let G be a planar graph. Then*

$$\chi(G^2) \leq \begin{cases} \Delta(G) + 5 & \text{if } 4 \leq \Delta(G) \leq 7, \\ \lfloor 3\Delta(G)/2 \rfloor + 1 & \text{if } \Delta(G) \geq 8. \end{cases}$$

*Received by the editors June 6, 2001; accepted for publication (in revised form) May 29, 2003; published electronically November 14, 2003. This work was supported by the Institute of Mathematics, Academia Sinica, Taipei.

http://www.siam.org/journals/sidma/17-2/39044.html

[†] Department of Mathematics, Zhejiang Normal University, Jinhua, Zhejiang 321004, People's Republic of China (wangweifan@mail.zjnu.net.cn). This work was done while this author was visiting the Institute of Mathematics, Academia Sinica, Taipei.

[‡] Institute of Mathematics, Academia Sinica, Nankang, Taipei 115, Taiwan (makwlih@sinica.edu.tw).

This conjecture remains open. Van den Heuvel and McGuinness [7] have proved $\chi(G^2) \le 2\Delta(G) + 25$ for any planar graph G. The main result in Borodin, Broersma, Glebov, and van den Heuvel [1] implies that, for a planar graph G, $\chi(G^2) \le \lceil \frac{9}{5}\Delta(G)\rceil + 1$ when $\Delta(G) \ge 47$. The present authors in [10] proved that every outerplanar graph G satisfies $\chi(G^2) \le \Delta(G) + 2$. Moreover, $\chi(G^2) = \Delta(G) + 1$ when $\Delta(G) \ge 7$. This establishes Conjecture 1 for outerplanar graphs.

Let p, q be two nonnegative integers. An *$L(p,q)$-labeling* of a graph G is a function ϕ from its vertex set $V(G)$ to the set $\{0, 1, \ldots, k\}$ for some positive integer k such that $|\phi(x) - \phi(y)| \ge p$ if x and y are adjacent, and $|\phi(x) - \phi(y)| \ge q$ if x and y are at distance 2. The *$L(p,q)$-labeling number* $\lambda(G; p, q)$ of G is the smallest k such that G has an $L(p,q)$-labeling with $\max\{\phi(v) \mid v \in V(G)\} = k$. An $L(1,1)$-labeling of the graph G is a proper coloring of its square G^2, and $\lambda(G; 1, 1) = \chi(G^2) - 1$. Georges and Mauro [4] obtained $\lambda(G; p, q)$ for cycles, paths, complete multipartite graphs, and t-point suspensions of paths and cycles.

The $L(2,1)$-labelings have been studied rather extensively in recent years [2, 3, 5, 6, 7, 8, 11, 14]. It is clear that $\lambda(G; 2, 1) \ge \Delta(G) + 1$ for any graph G. Griggs and Yeh [6] proposed the following conjecture.

Conjecture 2. *For any graph G with $\Delta(G) \ge 2$, we have $\lambda(G; 2, 1) \le \Delta^2(G)$.*

Conjecture 2 was verified in [6] for a few special cases, e.g., paths, cycles, trees, graphs with diameter 2, etc. The best known upper bound is $\lambda(G; 2, 1) \le \Delta^2(G) + \Delta(G)$ by Chang and Kuo [2]. It was proved in [7] that $\lambda(G; p, q) \le (4q-2)\Delta(G) + 10p + 38q - 23$ for every planar graph G. This result implies that $\lambda(G; 2, 1) \le 2\Delta(G) + 35$ for every planar graph G. It is reported in [1] that $\lambda(G; p, q) \le \lceil \frac{9}{5}\Delta(G)\rceil(2q - 1) + 8p - 8q + 1$ for every planar graph G with $\Delta(G) \ge 47$. This in turn implies that $\lambda(G; 2, 1) \le \lceil \frac{9}{5}\Delta(G)\rceil + 9$ for such graphs.

In this paper, we study the $L(p,q)$-labeling of a planar graph with conditions on its girth. More precisely, we will prove the following.

Theorem 1. *Let G be a planar graph and p and q two positive integers.*

(1) *If $g(G) \ge 7$, then $\lambda(G; p, q) \le (2q - 1)\Delta(G) + 4p + 4q - 4$.*

(2) *If $g(G) \ge 6$, then $\lambda(G; p, q) \le (2q - 1)\Delta(G) + 6p + 12q - 9$.*

(3) *If $g(G) \ge 5$, then $\lambda(G; p, q) \le (2q - 1)\Delta(G) + 6p + 24q - 15$.*

Corollary 2. *Let G be a planar graph.*

(1) *If $g(G) \ge 7$, then $\chi(G^2) \le \Delta(G) + 5$ and $\lambda(G; 2, 1) \le \Delta(G) + 8$.*

(2) *If $g(G) \ge 6$, then $\chi(G^2) \le \Delta(G) + 10$ and $\lambda(G; 2, 1) \le \Delta(G) + 15$.*

(3) *If $g(G) \ge 5$, then $\chi(G^2) \le \Delta(G) + 16$ and $\lambda(G; 2, 1) \le \Delta(G) + 21$.*

Note that $\lfloor 3\Delta(G)/2\rfloor + 1 \ge \Delta(G) + 5$ when $\Delta(G) \ge 8$, $\lfloor 3\Delta(G)/2\rfloor + 1 \ge \Delta(G) + 10$ when $\Delta(G) \ge 18$, and $\lfloor 3\Delta(G)/2\rfloor + 1 \ge \Delta(G) + 16$ when $\Delta(G) \ge 30$. Thus the following result is an immediate consequence of Corollary 2.

Corollary 3. *Conjecture* 1 *holds for planar graphs G with $g(G) \ge 7$, or $g(G) = 6$ and $\Delta(G) \ge 18$, or $g(G) = 5$ and $\Delta(G) \ge 30$.*

Furthermore, since $\Delta(G) + 8 \le \Delta^2(G)$ when $\Delta(G) \ge 4$, $\Delta(G) + 15 \le \Delta^2(G)$ when $\Delta(G) \ge 5$, and $\Delta(G) + 21 \le \Delta^2(G)$ when $\Delta(G) \ge 6$, we have the following.

Corollary 4. *Conjecture* 2 *holds for planar graphs G with $g(G) \ge 7$ and $\Delta(G) \ge 4$, or $g(G) = 6$ and $\Delta(G) \ge 5$, or $g(G) = 5$ and $\Delta(G) \ge 6$.*

2. Structural lemmas.

Let G be a plane graph. For $f \in F(G)$, we use $b(f)$ to denote the boundary walk of f and write $f = [u_1 u_2 \ldots u_n]$ if $u_1, u_2, \ldots, u_n$ are the vertices of $b(f)$ in the clockwise order. Repeated occurrences of a vertex are allowed. The degree of a face is the number of edge-steps in its boundary walk. Note that each cut-edge is counted twice. For $x \in V(G) \cup F(G)$, let $d_G(x)$, or simply $d(x)$, denote the degree of x in G. A vertex (or face) of degree k is called a *k-vertex* (or *k-face*). We say that f is an *$(m_1, m_2, \ldots, m_n)$-face* if $d(u_i) = m_i$ for $i = 1, 2, \ldots, n$. Let $V_i(f)$ denote the set of i-vertices incident to the face f. Let $p_i(f)$ denote the number of occurrences of i-vertices in $b(f)$. When v is a k-vertex, we say that there are k faces incident to v. However, these faces are not required to be distinct, i.e., v may have repeated occurrences on the boundary walk of some of its incident faces.

In this section, we always assume that G is a connected plane graph with minimum degree at least 2.

Using Euler's formula $|V(G)| - |E(G)| + |F(G)| = 2$ and $\sum\{d(v) \mid v \in V(G)\} = \sum\{d(f) \mid f \in F(G)\} = 2|E(G)|$, we can derive the following identity:

$$\sum_{v \in V(G)} (d(v) - 4) + \sum_{f \in F(G)} (d(f) - 4) = -8. \tag{1}$$

Lemmas 6, 7, and 8 of this section will be proved by the method of contradiction. In each case, we assume that G is a counterexample to the lemma under consideration. We define the weight function w by $w(x) = d(x) - 4$ for all $x \in V(G) \cup F(G)$. It follows from identity (1) that the total sum of weights is equal to -8. In each lemma, we will define appropriate discharging rules and redistribute weights accordingly. Once the discharging is finished, a new weight function w' is produced. However, the total sum of weights is kept fixed when the discharging is in process. Nevertheless, we can show that $w'(x) \geq 0$ for all $x \in V(G) \cup F(G)$. This leads to the following obvious contradiction,

$$0 \leq \sum\{w'(x) \mid x \in V(G) \cup F(G)\} = \sum\{w(x) \mid x \in V(G) \cup F(G)\} = -8 < 0,$$

and hence demonstrates that no such counterexample can exist.

For $x, y \in V(G) \cup F(G)$, we will use $\mathbf{W}(x \to y)$ to denote the sum of weights discharged from x to y and $\mathbf{W}(x \to)$ to denote the total weight discharged from x to all its adjacent or incident elements.

Lemma 5. *Let G be a plane graph such that no two adjacent vertices in G can both be 2-vertices. Let f be a face of G. Then $p_2(f) \leq \lfloor d(f)/2 \rfloor$. Furthermore, if, for some $k \geq 3$, G does not contain any path xyz such that $d(x) = 2$, $d(y) = 3$, and $d(z) \leq k$, then $p_3(f) \leq d(f) - 2p_2(f)$.*

Proof. The first conclusion is obvious. The second conclusion holds because, if we start moving along $b(f)$ from a 2-vertex, then we must encounter a vertex of degree at least 4 before we reach the next 2-vertex. □

Lemma 6. *Let G be a connected plane graph with $\delta(G) = 2$ and $g(G) \geq 7$. Then G contains a vertex v whose neighbors $v_1, v_2, \ldots, v_k$ satisfy one of the following conditions, assuming $d(v_1) \leq d(v_2) \leq \cdots \leq d(v_k)$:*

(A1) *$k = 2$ and $d(v_1) = 2$;*

(A2) *$k = 3$, $d(v_1) = 2$, and $d(v_2) \leq 3$;*

(A3) *$k = 3$, $d(v_1) = 2$, and $d(v_2) = d(v_3) = 4$ such that v_i is adjacent to a 2-vertex u_i for $i = 2$ and 3.*

(A4) *$k = 4$, either $d(v_1) = d(v_2) = d(v_3) = 2$ or $d(v_1) = d(v_2) = 2$ and $d(v_3) = 3$ such that v_3 is adjacent to a 2-vertex u_3.*

Proof. Suppose that the lemma is false. Let G be a connected plane graph with $\delta(G) = 2$ and $g(G) \geq 7$ such that none of its vertices satisfies one of (A1), (A2), (A3), and (A4). The discharging rules are defined as follows.

(R1) Let v be a vertex with degree at least 5 and f a face incident to v. For each occurrence of v in $b(f)$, we transfer the amount $w(v)/d(v)$ from v to f.

(R2) Let $\alpha(f)$ denote the sum of weights discharged into the face f from its incident vertices of degree at least 5 according to (R1). For each occurrence of a vertex v in $b(f)$, we transfer the amount 1 from f to v if $d(v) = 2$, or the amount $(w(f) + \alpha(f) - p_2(f))/p_3(f)$ from f to v if $d(v) = 3$.

We carry out (R1) and (R2) in succession. Let w' denote the resultant weight function after discharging.

Claim 1. If f is a face of degree at least 8 and v is a 3-vertex incident to f, then $\mathbf{W}(f \to v) \geq 1/2$.

Suppose $f = [u_1u_2 \ldots u_{d(f)}]$. Since (A1) and (A2) fail, G does not contain two adjacent 2-vertices and a path xyz with $d(x) = 2$, $d(y) = 3$, and $d(z) \leq 3$. Hence $p_2(f) \leq \lfloor d(f)/2 \rfloor$ and $p_3(f) \leq d(f) - 2p_2(f)$ by Lemma 5. If $d(f) \geq 8$, then $2(w(f) + \alpha(f) - p_2(f)) \geq 2(w(f) - p_2(f)) \geq 2d(f) - 8 - d(f) + p_3(f) \geq p_3(f)$. This proves that $\mathbf{W}(f \to v) \geq 1/2$.

Claim 2. Suppose that f is a 7-face and v is a 3-vertex incident to f. Then one of the following holds.

(1) $\mathbf{W}(f \to v) \geq 1/2$ when $p_2(f) = 1$, or $p_2(f) = 0$ and $p_3(f) \leq 6$, or $p_2(f) = 2$ and $p_3(f) \leq 2$.

(2) $\mathbf{W}(f \to v) = 0$ when f is a $(2, 4, 2, 4, 2, 4, 3)$-face.

(3) $\mathbf{W}(f \to v) = 1/5$ when f is a $(2, 5, 2, 4, 2, 4, 3)$-face, or a $(2, 4, 2, 5, 2, 4, 3)$-face, or a $(2, 4, 2, 4, 2, 5, 3)$-face.

(4) $\mathbf{W}(f \to v) \geq 1/3$ in all other cases.

First note that $w(f) = 3$ and $p_2(f) \leq 3$. If $p_2(f) = 0$, then $\mathbf{W}(f \to v) = 3/7$ when $p_3(f) = 7$, and $\mathbf{W}(f \to v) \geq 1/2$ when $p_3(f) \leq 6$. If $p_2(f) = 1$, then $p_3(f) \leq 4$, and hence $\mathbf{W}(f \to v) \geq 1/2$. If $p_2(f) = 2$, then $p_3(f) \leq 3$. When $p_3(f) = 3$, $\mathbf{W}(f \to v) \geq (w(f) + \alpha(f) - p_2(f))/3 \geq 1/3$. When $p_3(f) \leq 2$, we have $\mathbf{W}(f \to v) \geq 1/2$.

Now assume that $p_2(f) = 3$. Then $p_3(f) \leq 1$ and v is the only 3-vertex on $b(f)$. If $b(f)$ does not contain a vertex of degree at least 5, then f is a $(2, 4, 2, 4, 2, 4, 3)$-face and $\mathbf{W}(f \to v) = 0$. If $b(f)$ contains at least one vertex, say x, of degree ≥ 6, then we transfer at least 1/3 from x to f according to (R1), and $\alpha(f) \geq 1/3$. Thus $\mathbf{W}(f \to v) \geq w(f) + \alpha(f) - p_2(f) \geq 1/3$. If $b(f)$ contains at least two 5-vertices, say x_1 and x_2, then $\mathbf{W}(f \to v) \geq w(f) + \alpha(f) - p_2(f) = \alpha(f) \geq \mathbf{W}(x_1 \to f) + \mathbf{W}(x_2 \to f) \geq 2/5$ according to (R1). If $b(f)$ contains exactly one 5-vertex, we have $\mathbf{W}(f \to v) = 1/5$. In this case, f is a $(2, 5, 2, 4, 2, 4, 3)$-face, or a $(2, 4, 2, 5, 2, 4, 3)$-face, or a $(2, 4, 2, 4, 2, 5, 3)$-face. This concludes the proof of Claim 2.

It remains to verify that $w'(x) \geq 0$ for all $x \in V(G) \cup F(G)$. Let $f \in F(G)$. Since the degree of a face is never less than the girth of the graph, we have $d(f) \geq 7$. Since $p_2(f) \leq \lfloor d(f)/2 \rfloor$ and $w(f) + \alpha(f) - p_2(f) \geq d(f) - 4 - \lfloor d(f)/2 \rfloor = \lceil d(f)/2 \rceil - 4 \geq 0$, it follows that $w'(f) \geq 0$ by (R2).

Next let $v \in V(G)$. If $d(v) = 2$, we transfer 1 into v from each of the incident faces of v by (R2). Thus $w'(v) = w(v) + 2 = 0$. If $d(v) = 4$, then $w'(v) = w(v) = 0$. If $d(v) \geq 5$, then $w(v) = d(v) - 4 \geq 1$, and (R1) implies that $w'(v) = 0$.

Finally, assume that $d(v) = 3$. Thus $w(v) = -1$. Let f_1, f_2, and f_3 denote the faces in G incident to v. We may assume that, among the three faces, f_1 discharges the least amount to v. If $\mathbf{W}(f_1 \to v) \geq 1/3$, then $w'(v) \geq 0$. Otherwise, according to Claim 2, we only need to consider the following two cases.

Case 1. $\mathbf{W}(f_1 \to v) = 0$.

From Claims 1 and 2, we know that f_1 is a $(2, 4, 2, 4, 2, 4, 3)$-face. We will show that $\mathbf{W}(f_i \to v) \geq 1/2$ for $i = 2, 3$, and thus $w'(v) \geq 0$. Let $f_1 = [vxx_1x_2x_3x_4y]$ with $d(x) = d(x_2) = d(x_4) = 2$ and $d(x_1) = d(x_3) = d(y) = 4$. We may suppose that f_1

and f_2 share the edge vy and f_1 and f_3 share the path vxx_1.

By Claim 1, $\mathbf{W}(f_2 \to v) \geq 1/2$ if $d(f_2) \geq 8$. Now let $f_2 = [vyy_1y_2y_3y_4z]$. Since (A2) fails at v, we must have $d(z) \geq 4$. Since (A4) fails at y, we also have $d(y_1) \geq 3$. If $p_2(f_2) = 0$, then $p_3(f_2) \leq 5$, and hence $\mathbf{W}(f_2 \to v) \geq 3/5$. If exactly one among y_2, y_3, and y_4 is a 2-vertex, then $\mathbf{W}(f_2 \to v) \geq 1/2$. If exactly two of them are 2-vertices, we must have $d(y_2) = d(y_4) = 2$. Thus $d(y_3) \geq 4$. It follows that $p_3(f_2) \leq 2$ and $\mathbf{W}(f_2 \to v) \geq 1/2$.

Similarly, we have $\mathbf{W}(f_3 \to v) \geq 1/2$ if $d(f_3) \geq 8$. Now let $f_3 = [vzz_1z_2z_3x_1x]$. Since (A4) fails at x_1, we have $d(z_3) \geq 3$.

Suppose that $d(z_3) = 3$. Again since x_1 does not satisfy (A4), we have $d(z_2) \geq 3$. If $d(z_2) = 3$, then $d(z_1) \geq 3$, implying $p_2(f_3) = 1$, and hence $\mathbf{W}(f_3 \to v) \geq 1/2$ by Claim 2. If $d(z_2) \geq 4$, then either $p_2(f_3) = 1$ or $p_2(f_3) = 2$ and $p_3(f_3) \leq 2$, hence $\mathbf{W}(f_3 \to v) \geq 1/2$.

Suppose that $d(z_3) \geq 4$. In this case, either $p_2(f_3) = 1$, or $p_2(f_3) = 2$ and $p_3(f_3) \leq 2$. Therefore $\mathbf{W}(f_3 \to v) \geq 1/2$.

Case 2. $\mathbf{W}(f_1 \to v) = 1/5$.

Again let $f_1 = [vxx_1x_2x_3x_4y]$ such that f_1 and f_2 share the edge vy, and f_1 and f_3 share the path vxx_1. By Claim 2, we first suppose that $d(x) = d(x_2) = d(x_4) = 2$, $d(x_3) = d(y) = 4$, and $d(x_1) = 5$, i.e., f_1 is a $(2,5,2,4,2,4,3)$-face.

Since (A2) fails at v, we have $d(z) \geq 4$. Similar to Case 1, $\mathbf{W}(f_2 \to v) \geq 1/2$. Let $f_3 = [vzz_1z_2z_3x_1x]$. Since the degrees of the three consecutive vertices x_1, x, v are 5, 2, 3, f_3 cannot be a $(2,4,2,4,2,4,3)$-face, or a $(2,4,2,5,2,4,3)$-face, or a $(2,4,2,4,2,5,3)$-face. If it is a $(4,2,4,2,5,2,3)$-face, then $d(z_1) = d(z_3) = 2$ and $d(z) = d(z_2) = 4$. Thus, v satisfies (A3) since y and z are, respectively, adjacent to at least one 2-vertex. This is not allowed. From Claim 2, all remaining possibilities lead to $\mathbf{W}(f_3 \to v) \geq 1/3$. Consequently, $w'(v) \geq -1 + (1/5) + (1/2) + (1/3) = 1/30$.

Similar arguments can be constructed when f_1 is either a $(2,4,2,5,2,4,3)$-face or a $(2,4,2,4,2,5,3)$-face. This concludes the proof of Case 2. □

LEMMA 7. *Let G be a connected plane graph with $\delta(G) = 2$ and $g(G) \geq 6$. If G does not contain a 6-face $[u_1u_2 \ldots u_6]$ satisfying $d(u_2) \leq 5$ and $d(u_1) = d(u_3) = d(u_5) = 2$, then G contains a vertex v whose neighbors $v_1, v_2, \ldots, v_k$ satisfy one of the following conditions, assuming $d(v_1) \leq d(v_2) \leq \cdots \leq d(v_k)$:*

(B1) *$k = 2$ and $d(v_1) = 2$;*

(B2) *$k = 3$, $d(v_1) = 2$, and $d(v_2) \leq 5$;*

(B3) *$k = 4$, $d(v_1) = 2$, $d(v_2) \leq 3$, and $d(v_3) \leq 5$.*

Proof. Suppose that the lemma is false. Let G be a connected plane graph with $\delta(G) = 2$, $g(G) \geq 6$, and without any 6-face $[u_1u_2 \ldots u_6]$ satisfying $d(u_2) \leq 5$ and $d(u_1) = d(u_3) = d(u_5) = 2$. We further assume that none of its vertices satisfies one of (B1), (B2), and (B3).

For $f \in F(G)$, let $V_3'(f)$ denote the set of those 3-vertices on the boundary of f, each of which is not adjacent to any 2-vertex on that boundary. Thus $V_3'(f) \subseteq V_3(f)$. Furthermore, for $k = 0, 1, 2$, let $V_2^k(f)$ denote the set of those 2-vertices on the boundary of f, each of which is adjacent to k vertices of degree at least 6 on that boundary. By definition, $V_2(f) = V_2^0(f) \cup V_2^1(f) \cup V_2^2(f)$.

We define the following discharging rules.

(R1) For every vertex v of degree at least 6, we transfer the amount $w(v)/d(v)$ from v to each of its adjacent vertices of degree less than 6.

(R2) For a vertex v with $4 \leq d(v) \leq 5$, let $\gamma(v)$ denote the sum of weights discharged to v from its adjacent vertices of degree at least 6 according to (R1). If f

is a face incident to v, then, for each occurrence of v in $b(f)$, we transfer the amount $(w(v)+\gamma(v))/d(v)$ from v to f.

(R3) Let v be a vertex of degree at most 3 and f a face incident to v. For each occurrence of v in $b(f)$, we transfer the amount 1/3 from f to v when $v \in V_3'(f)$, or 2/3 from f to v when $v \in V_2^2(f)$, or 5/6 from f to v when $v \in V_2^1(f)$, or 1 from f to v when $v \in V_2^0(f)$.

We carry out (R1), (R2), and (R3) in succession. Let w' denote the resultant weight function after discharging. It remains to verify that $w'(x) \geq 0$ for all $x \in V(G) \cup F(G)$.

Let $v \in V(G)$. If $d(v) \geq 6$, then $w(v) \geq 2$ and $w'(v) \geq 0$ by (R1). If $4 \leq d(v) \leq 5$, then $w(v)+\gamma(v) \geq w(v) \geq 0$ and thus $w'(v) \geq 0$ by (R2). Assume that $d(v) = 3$. Thus $w(v) = -1$. Let x_1, x_2, and x_3 denote the neighbors of v in G with $d(x_1) \leq d(x_2) \leq d(x_3)$. If $d(x_1) \geq 3$, then v receives three times 1/3 from the incident faces by (R3), and $w'(v) \geq 0$. If $d(x_1) = 2$, then $d(x_2) \geq 6$ and $d(x_3) \geq 6$ since (B2) fails at v. It follows from (R1) that, for $i = 2$ and 3, $\mathbf{W}(x_i \to v) = (d(x_i)-4)/d(x_i) = 1 - 4/d(x_i) \geq 1/3$. Note that the face whose boundary walk contains the path x_3vx_2 discharges 1/3 to v by (R3). Hence $w'(v) \geq 0$.

Assume that $d(v) = 2$. So $w(v) = -2$. Let y_1 and y_2 denote the neighbors of v with $d(y_1) \leq d(y_2)$, and f_1 and f_2 denote the faces of G incident to v. If $d(y_1) \geq 6$, then for $i = 1$ and 2, $\mathbf{W}(y_i \to v) \geq 1/3$ by (R1) and $\mathbf{W}(f_i \to v) = 2/3$ by (R3). Consequently, $w'(v) \geq -2 + 2(2/3) + 2(1/3) = 0$. If $d(y_1) \leq 5$, we consider two possibilities. When $d(y_2) \leq 5$, each of f_1 and f_2 discharges 1 to v by (R3), and hence $w'(v) \geq -2+2 = 0$. When $d(y_2) \geq 6$, $\mathbf{W}(y_2 \to v) \geq 1/3$ and $\mathbf{W}(f_1 \to v) = \mathbf{W}(f_2 \to v) = 5/6$, hence $w'(v) \geq -2 + 2(5/6) + (1/3) = 0$.

Let $f \in F(G)$ and let $\alpha(f)$ denote the sum of weights discharged into f from all its incident 4-vertices and 5-vertices according to (R2). Suppose that $f = [u_1u_2 \ldots u_{d(f)}]$. Since G does not contain two adjacent 2-vertices, $p_2(f) \leq \lfloor d(f)/2 \rfloor$ by Lemma 5. By (R3), $\mathbf{W}(f \to) = (|V_3'(f)|/3) + (2|V_2^2(f)|/3) + (5|V_2^1(f)|/6) + |V_2^0(f)|$. We will show $\mathbf{W}(f \to) \leq w(f) + \alpha(f)$. Then $w'(f) = w(f) + \alpha(f) - \mathbf{W}(f \to) \geq 0$. It is obvious that $\mathbf{W}(f \to) \leq p_2(f) + p_3(f)/3$.

Let $d(f) \geq 8$. Since G does not contain a path xyz with $d(x) = 2$, $d(y) = 3$, and $d(z) \leq 5$, it follows that $p_3(f) \leq d(f) - 2p_2(f)$ by Lemma 5. Thus $\mathbf{W}(f \to) \leq p_2(f)+p_3(f)/3 \leq p_2(f)+(d(f)-2p_2(f))/3 = (d(f)+p_2(f))/3 \leq d(f)/2 \leq d(f)-4 = w(f) \leq w(f)+\alpha(f)$.

Let $d(f) = 7$. Then $w(f) = 3$ and $p_2(f) \leq 3$. If $p_2(f) \leq 1$, then $\mathbf{W}(f \to) \leq 1+6(1/3) = 3$. If $p_2(f) = 2$, then $p_3(f) \leq 3$, and $\mathbf{W}(f \to) \leq 2+3(1/3) = 3$. Assume that $p_2(f) = 3$. So $p_3(f) \leq 1$. If $p_3(f) = 0$, then $\mathbf{W}(f \to) \leq 3$. If $p_3(f) = 1$, then the unique 3-vertex on the boundary of f, say z, is adjacent to some 2-vertex on the boundary. By (R3), nothing is discharged from f to z. Thus $\mathbf{W}(f \to) \leq 3$.

Let $d(f) = 6$. Thus $w(f) = 2$ and $p_2(f) \leq 3$. Suppose $d(u_1) = d(u_3) = d(u_5) = 2$. By our assumptions on G, the degrees of u_2, u_4, and u_6 must all be greater than 5. Now we transfer 2/3 from f to each incident 2-vertex by (R3), and hence $\mathbf{W}(f \to) = 2$. If $p_2(f) = 0$, it is clear that $\mathbf{W}(f \to) \leq 6(1/3) = 2$. If $p_2(f) = 1$, it is easy to see that $p_3(f) \leq 3$ and $\mathbf{W}(f \to) \leq 1+3(1/3) = 2$. Finally, suppose that $p_2(f) = 2$. It suffices to handle the following two cases.

Case 1. $d(u_1) = d(u_4) = 2$.

Note that any 3-vertex on $b(f)$, if it exists, must be adjacent to either u_1 or u_4. It follows from (R3) that $\mathbf{W}(f \to) \leq 2$.

Case 2. $d(u_1) = d(u_3) = 2$.

It is easy to see that $d(u_2) \geq 4$ since (B2) fails at u_2. Moreover, $p_3(f) \leq 2$.

If $p_3(f) = 0$, then $\mathbf{W}(f \rightarrow) \leq 2$.

Assume that $p_3(f) = 1$. If either $d(u_4) = 3$ or $d(u_6) = 3$, then it is obvious that $\mathbf{W}(f \rightarrow) \leq 2$. Suppose $d(u_5) = 3$. We see that $d(u_4) \geq 4$ and $d(u_6) \geq 4$ as (B2) fails at these vertices, and $\mathbf{W}(f \rightarrow u_5) = 1/3$ by (R3).

If $d(u_2) \geq 6$, then $\mathbf{W}(f \rightarrow u_i) \leq 5/6$ for $i = 1$ and 3. Thus $\mathbf{W}(f \rightarrow) \leq 2(5/6) + 1/3 = 2$. If $d(u_2) = 5$, then $\mathbf{W}(u_2 \rightarrow f) = (w(u_2) + \gamma(u_2))/d(u_2) \geq (d(u_2) - 4)/d(u_2) = 1/5$. If $d(u_2) = 4$, let z_1 and z_2 denote the other two neighbors of u_2 different from u_1 and u_3. For $i = 1$ and 2, we have $d(z_i) \geq 6$ since (B3) fails at u_2, and $\mathbf{W}(z_i \rightarrow u_2) \geq 1/3$ by (R1). So $\gamma(u_2) \geq 2/3$ and $\mathbf{W}(u_2 \rightarrow f) \geq \gamma(u_2)/d(u_2) \geq 1/6$ by (R2).

In summary, u_2 discharges at least 1/6 to the face f when $4 \leq d(u_2) \leq 5$. This implies that $\alpha(f) \geq 1/6$. If $d(u_4) \geq 6$, then $\mathbf{W}(f \rightarrow u_3) = 5/6$ by (R3), and hence $\mathbf{W}(f \rightarrow) \leq (5/6) + 1 + (1/3) = 2 + 1/6 \leq w(f) + \alpha(f)$. If $4 \leq d(u_4) \leq 5$, we can reason similarly to obtain $\mathbf{W}(u_4 \rightarrow f) \geq 1/6$. Thus $\alpha(f) \geq \mathbf{W}(u_2 \rightarrow f) + \mathbf{W}(u_4 \rightarrow f) \geq 1/3$. Hence $\mathbf{W}(f \rightarrow) \leq 2 + 1/3 \leq w(f) + \alpha(f)$.

Finally, assume $p_3(f) = 2$. The only possibility is $d(u_4) = d(u_6) = 3$. Since both u_4 and u_6 are adjacent to a 2-vertex on the boundary of f, it follows that $\mathbf{W}(f \rightarrow) \leq 2$ by (R3). □

LEMMA 8. *Let G be a connected plane graph with $g(G) \geq 5$, $\delta(G) \geq 2$, and without a path $x_1x_2x_3x_4$ such that $d(x_2) = d(x_3) = 3$, $d(x_1) \leq 11$, and $d(x_4) \leq 11$. Then G contains a vertex v whose neighbors $v_1, v_2, \ldots, v_k$ satisfy one of the following conditions, assuming $d(v_1) \leq d(v_2) \leq \cdots \leq d(v_k)$:*

(C1) *$k = 2$ and $d(v_1) = 2$;*

(C2) *$k = 3$, $d(v_1) = 2$, and $d(v_2) \leq 11$;*

(C3) *$k = 4$, $d(v_1) = 2$, $d(v_2) \leq 7$, and $d(v_3) \leq 7$;*

(C4) *$k = 5$, $d(v_1) = d(v_2) = d(v_3) = 2$, and $d(v_4) \leq 7$;*

(C5) *$k = 6$ and $d(v_1) = d(v_2) = d(v_3) = d(v_4) = d(v_5) = 2$;*

(C6) *$k = 7$ and $d(v_1) = d(v_2) = \cdots = d(v_7) = 2$.*

Proof. Suppose that the lemma is false. Let G be a connected plane graph with $g(G) \geq 5$, $\delta(G) \geq 2$, and without any path $x_1x_2x_3x_4$ satisfying $d(x_2) = d(x_3) = 3$, $d(x_1) \leq 11$, and $d(x_4) \leq 11$. We further assume that none of its vertices satisfies one of (C1) to (C6).

For a face $f \in F(G)$, let $V_3^*(f)$ denote the set of 3-vertices in $V_3(f)$, each of which has two neighbors of degree at most 7 but no neighbors of degree 2. Let $V_2^*(f)$ denote the set of 2-vertices in $V_2(f)$, each of which has two neighbors of degree at least 12.

The discharging rules are defined as follows:

(R1) For every vertex v of degree at least 8, we transfer the amount $w(v)/d(v)$ from v to each of its adjacent vertices of degree at most 7.

(R2) For a vertex v with $3 \leq d(v) \leq 7$, we transfer the amount 1/2 from v to each of its adjacent 2-vertices.

(R3) Let f be a face of degree at least 6 incident to a vertex v. For each occurrence of v in $b(f)$, we transfer the amount 1/2 from f to v if $v \in V_2(f)$ and the amount 1/3 from f to v if $v \in V_3(f)$.

(R4) Let f be a 5-face. For each occurrence of a vertex v in $b(f)$, we transfer the amount 1/3 from f to v if $v \in V_2^*(f) \cup V_3^*(f)$ and the amount 1/2 from f to v if $v \in V_2(f) \setminus V_2^*(f)$. Afterwards, the remaining weight of f is evenly distributed to other 3-vertices in $b(f)$.

We carry out (R1), (R2), (R3), and (R4) in succession. Let w' denote the resultant weight function after discharging. It remains to verify that $w'(x) \geq 0$ for all $x \in V(G) \cup F(G)$.

Let $f \in F(G)$. Since $g(G) \geq 5$, we have $d(f) \geq 5$. Again, we know that $p_2(f) \leq \lfloor d(f)/2 \rfloor$ by Lemma 5. It suffices to show $\mathbf{W}(f \rightarrow) \leq w(f)$. If $d(f) \geq 7$, it follows from (R3) that $\mathbf{W}(f \rightarrow) \leq p_2(f)/2 + p_3(f)/3 \leq p_2(f)/2 + (d(f) - p_2(f))/3 = d(f)/3 + p_2(f)/6 \leq d(f)/3 + d(f)/12 = 5d(f)/12 \leq d(f) - 4 = w(f)$.

Assume that $d(f) = 6$. We see that $w(f) = 2$ and $p_2(f) \leq 3$. If $p_2(f) = 3$, then $p_3(f) = 0$ and $\mathbf{W}(f \rightarrow) \leq 3/2$. If $1 \leq p_2(f) \leq 2$, then $p_3(f) \leq 3$ and $\mathbf{W}(f \rightarrow) \leq 2(1/2) + 3(1/3) = 2$. If $p_2(f) = 0$, then $\mathbf{W}(f \rightarrow) \leq 6(1/3) = 2$.

Assume that $d(f) = 5$. Then $w(f) = 1$ and $p_2(f) \leq 2$. Let $\beta(f) = (|V_2^*(f)| + |V_3^*(f)|)/3 + |V_2(f) \setminus V_2^*(f)|/2$. It suffices to prove $\beta(f) \leq 1$. If $p_2(f) = 0$, we claim that $p_3(f) \leq 3$ and thus $\beta(f) \leq 3(1/3) = 1$. In fact, if $p_3(f) \geq 4$, then G would contain a path $x_1x_2x_3x_4$ such that $d(x_i) = 3$ for all $1 \leq i \leq 4$, contradicting the assumptions on G. Let $p_2(f) = 1$. Suppose that $f = [u_1u_2u_3u_4u_5]$ and $d(u_1) = 2$. Both u_2 and u_5 cannot belong to $V_3^*(f)$. If at most one of u_3 and u_4 is of degree 3, then $\beta(f) \leq 1/2 + 1/3 = 5/6$ by (R4). Assume that $d(u_3) = d(u_4) = 3$. If both u_2 and u_5 are of degree at least 12, then $\mathbf{W}(f \rightarrow u_1) = 1/3$ by (R4), and thus $\beta(f) \leq 3(1/3) = 1$. Suppose that at least one of u_2 and u_5, say u_2, is of degree at most 11. Let z denote the neighbor of u_4 that differs from u_3 and u_5. It follows that $d(z) \geq 12$ and $d(u_5) \geq 12$, for otherwise $u_2u_3u_4z$ or $u_2u_3u_4u_5$ would be a forbidden path. Thus $u_4 \notin V_3^*(f)$ and $\beta(f) \leq 5/6$. Let $p_2(f) = 2$. It is easy to see that each 3-vertex on $b(f)$, if it exists, is adjacent to some 2-vertex on $b(f)$. Hence $V_3^*(f) = \emptyset$ and $\beta(f) \leq 1$.

Let $v \in V(G)$. List all neighbors of v as $v_1, v_2, \ldots, v_k$ such that $d(v_1) \leq d(v_2) \leq \cdots \leq d(v_k)$. If $d(v) \geq 8$, it is obvious that $w'(v) \geq 0$ by (R1). If $d(v) = 2$, then $w(v) = -2$. Note that $d(v_1) \geq 3$. Let f_1 and f_2 denote the faces incident to v. When $d(v_1) \geq 12$, $\mathbf{W}(v_1 \rightarrow v) \geq (d(v_1) - 4)/d(v_1) \geq 2/3$, $\mathbf{W}(v_2 \rightarrow v) \geq 2/3$, and $\mathbf{W}(f_i \rightarrow v) = 1/3$ for $i = 1$ and 2 by (R1), (R3), and (R4). Consequently, $w'(v) \geq -2 + 2(2/3) + 2(1/3) = 0$. When $d(v_1) \leq 11$, each of v_1, v_2, f_1, and f_2 discharges at least 1/2 to v, thus $w'(v) \geq -2 + 2 = 0$.

Let $d(v) = 3$. If $d(v_1) = 2$, then $d(v_2) \geq 12$ and $d(v_3) \geq 12$ since (C2) fails at v. Let f^* denote the face of G whose boundary contains the path v_2vv_3. If $d(f^*) \geq 6$, then f^* discharges the amount 1/3 to v by (R3). Thus $w'(v) \geq -1 + 2(2/3) + 1/3 - 1/2 = 1/6$ by (R1) and (R2). Assume that $d(f^*) = 5$. Since both $d(v_2)$ and $d(v_3)$ are ≥ 12, we see that $p_2(f^*) \leq 1$ and $p_2(f^*) + p_3(f^*) \leq 3$. It follows from (R4) that $\mathbf{W}(f^* \rightarrow v) \geq 1/4$. Therefore $w'(v) \geq -1 + 2(2/3) + 1/4 - 1/2 = 1/12$. Suppose that $d(v_1) \geq 3$. If $d(v_2) \geq 8$, then $\mathbf{W}(v_i \rightarrow v) \geq 1/2$ for $i = 2$ and 3, thus $w'(v) \geq -1 + 2(1/2) = 0$. If $d(v_2) \leq 7$, then each of the faces incident to v discharges 1/3 to v by (R3) and (R4). Consequently, $w'(v) \geq -1 + 3(1/3) = 0$.

Let $d(v) = 4$. Thus $w(v) = 0$. If $d(v_1) \geq 3$, then it is evident that $w'(v) \geq 0$. If $d(v_1) = 2$, then $d(v_3) \geq 8$ since (C3) fails at v. By (R1), $\mathbf{W}(v_i \rightarrow v) \geq 1/2$ for $i = 3$ and 4. On the other hand, v discharges at most 1/2 to each of v_1 and v_2 by (R2). It follows that $w'(v) \geq 0$.

Let $d(v) = 5$. Thus $w(v) = 1$. If $d(v_3) \geq 3$, then $w'(v) \geq w(v) - 2(1/2) = 0$. If $d(v_3) = 2$, we see that $d(v_4) \geq 8$ since (C4) fails at v. So $w'(v) \geq 1 - 3(1/2) + 2(1/2) = 1/2$.

Let $d(v) = 6$. Since (C5) fails at v, it follows that $d(v_5) \geq 3$ and $w'(v) \geq w(v) - 4(1/2) = 0$.

Let $d(v) = 7$. Since (C6) fails at v, it follows that $d(v_7) \geq 3$ and $w'(v) \geq w(v) - 6(1/2) = 0$. □

3. Proof of the main theorem. Let G be a plane graph. For a vertex v of G and an integer $i \geq 1$, let $N_i(v)$ denote the set of vertices in G at distance i to v. Let $N_1(v) = \{x_1, x_2, \ldots, x_{d(v)}\}$. Suppose that we are trying to construct an $L(p,q)$-labeling ϕ of G and v is a yet-to-be-labeled vertex. For every labeled vertex $x \in N_1(v)$, there are $2p-1$ consecutive labels $\phi(x)-p+1, \phi(x)-p+2, \ldots, \phi(x), \phi(x)+1, \ldots, \phi(x)+p-1$ that are forbidden for use on v. Similarly, for every labeled vertex $y \in N_2(v)$, there are $2q-1$ consecutive labels $\phi(y)-q+1, \phi(y)-q+2, \ldots, \phi(y), \phi(y)+1, \ldots, \phi(y)+q-1$ that are forbidden for use on v. Let $\sigma(v)$ denote the number of labels forbidden for v. Then $\sigma(v) \leq (2p-1)d^*(v) + (2q-1)\sum\{d^*(x_i) \mid 1 \leq i \leq d(v)\}$, where $d^*(w)$ denotes the number of vertices adjacent to w that have already been labeled.

We are going to prove the following slightly stronger form of Theorem 1.

THEOREM 9. *For positive integers p, q, and M, let G be a plane graph satisfying $2 \leq \Delta(G) \leq M$. Then the following statements hold:*

(1) *If $g(G) \geq 7$, then $\lambda(G;p,q) \leq (2q-1)M + 4p + 4q - 4$.*

(2) *If $g(G) \geq 6$, then $\lambda(G;p,q) \leq (2q-1)M + 6p + 12q - 9$.*

(3) *If $g(G) \geq 5$, then $\lambda(G;p,q) \leq (2q-1)M + 6p + 24q - 15$.*

Proof. We may assume that G is connected and prove the theorem by induction on $|V(G)|+|E(G)|$. When $\Delta(G) = 2$, the theorem can be checked easily or it follows from results in [4]. Hence the induction basis holds for a cycle of length 5. Now let G be a plane graph with $|V(G)| + |E(G)| \geq 10$ satisfying $3 \leq \Delta(G) \leq M$. If there is a vertex v of degree 1, we can extend an $L(p,q)$-labeling of $G-v$ to an $L(p,q)$-labeling of G for cases (1), (2), and (3) since $\sigma(v) \leq (2p-1)+(2q-1)(\Delta(G)-1) \leq (2q-1)M+2p-2q$. Suppose that $\delta(G) \geq 2$.

Part 1. By Lemma 6, there is a vertex v with neighbors $v_1, v_2, \ldots, v_k$ such that at least one among (A1) to (A4) holds. Let $H = G - v_1$. Evidently, H is a plane graph with $g(H) \geq 7$ and $\Delta(H) \leq M$. By the induction hypothesis, H has an $L(p,q)$-labeling ϕ with the label set $L_1 = \{0, 1, \ldots, n_1\}$, where $n_1 = (2q-1)M + 4p + 4q - 4$. Now we label G as follows. Note that the order of labeling is essential for determining the bounds for forbidden labels.

If either (A1) or (A2) holds, erase $\phi(v)$. First label v, then label v_1.

If (A3) holds, erase $\phi(v), \phi(u_2)$, and $\phi(u_3)$. Then label v, u_2, u_3, and v_1 in succession.

If (A4) holds, erase $\phi(v), \phi(v_2)$, and $\phi(v_3)$ when $d(v_3) = 2$, or erase $\phi(v), \phi(v_2)$, and $\phi(u_3)$ when $d(v_3) = 3$. Afterwards, label v, v_2, v_3 (or u_3), and v_1 in succession.

For any vertex x that is ready to be labeled in the above process, the number of forbidden labels $\sigma(x)$ can be estimated as follows:

$$\begin{aligned}
\sigma(x) &\leq 2(2p-1) + (2q-1)(\Delta(G) - 1 + 3) \\
&\leq 2(2p-1) + (2q-1)(M+2) = n_1 \\
&\quad \text{for} \quad x \in \{v_1, v_2, v_3, u_2, u_3\}, \\
\sigma(v) &\leq \max\{2(2p-1) + 5(2q-1), 2(2p-1) + (2q-1)(\Delta(G)+2)\} \\
&\leq 2(2p-1) + (2q-1)(M+2) = n_1.
\end{aligned}$$

Thus ϕ can be extended to an $L(p,q)$-labeling of G with the label set L_1.

Part 2. If G contains a 6-face $[u_1 u_2 \ldots u_6]$ such that $d(u_i) = 2$ for $i = 1, 3, 5$, $d(u_2) \leq 5$, and $d(u_j) \leq \Delta(G)$ for $j = 4, 6$, then let $H = G - u_1 - u_3 + u_2 u_5$. Otherwise, there is a vertex v with neighbors $v_1, v_2, \ldots, v_k$ such that at least one among (B1) to (B3) in Lemma 7 holds. Let $H = G - v_1$. Obviously, H is a plane graph with $g(H) \geq 6$, $\Delta(H) \leq M$, and $|V(H)| + |E(H)| < |V(G)| + |E(G)|$. By the induction hypothesis, H has an $L(p,q)$-labeling ϕ with the label set $L_2 = \{0, 1, \ldots, n_2\}$, where $n_2 = (2q-1)M + 6p + 12q - 9$. Now we label G as follows.

If G contains a 6-face satisfying the conditions given above, we label u_1 and u_3 in succession.

If one of (B1), (B2), and (B3) holds, first erase $\phi(v)$. Then label v and v_1 in succession.

It is easy to see the following:

$$\begin{aligned}
\sigma(u_j) &\leq 2(2p-1) + (2q-1)(\Delta(G) - 1 + 4)\\
&\leq 2(2p-1) + (2q-1)(M+3)\\
&= (2q-1)M + 4p + 6q - 5 < n_2\\
&\quad \text{for} \quad j = 1, 3,\\
\sigma(v_1) &\leq 2(2p-1) + (2q-1)(\Delta(G) - 1 + 3)\\
&\leq (2q-1)M + 4p + 4q - 4 < n_2\\
\sigma(v) &\leq 3(2p-1) + (2q-1)(\Delta(G) - 1 + 1 + 2 + 4)\\
&\leq (2q-1)M + 6p + 12q - 9 = n_2.
\end{aligned}$$

Thus ϕ can be extended to an $L(p,q)$-labeling of G with the label set L_2.

Part 3. If G contains a path $x_1x_2x_3x_4$ with $d(x_2) = d(x_3) = 3$ and $d(x_i) \leq 11$ for $i = 1$ and 4, then let $H = G - x_2x_3$. Otherwise, there is a vertex v with neighbors $v_1, v_2, \ldots, v_k$ such that at least one among (C1) to (C6) in Lemma 8 holds. Let $H = G - v_1$. Obviously, H is a plane graph with $g(H) \geq 5$, $\Delta(H) \leq M$, and $|E(H)| < |E(G)|$. By the induction hypothesis, H has an $L(p,q)$-labeling ϕ with the label set $L_3 = \{0, 1, \ldots, n_3\}$, where $n_3 = (2q-1)M + 6p + 24q - 15$. Now we label G as follows.

If G contains a path satisfying the conditions given above, first erase $\phi(x_2)$ and $\phi(x_3)$. Then label x_2 and x_3 in succession. Note that $\sigma(x_i) \leq 3(2p-1) + (2q-1)(\Delta(G) - 1 + 10 + 2) \leq (2q-1)M + 6p + 22q - 14 < n_3$ for $i = 2$ and 3.

If one of (C1), (C2), and (C3) holds, first erase $\phi(v)$. Then label v and v_1 in succession. It is easy to see the following.

$$\begin{aligned}
\sigma(v_1) &\leq 2(2p-1) + (2q-1)(M+2) < n_3,\\
\sigma(v) &\leq \max\{(2p-1) + (2q-1)(\Delta(G) - 1 + 1),\\
&\qquad 2(2p-1) + (2q-1)(\Delta(G) - 1 + 10 + 1),\\
&\qquad 3(2p-1) + (2q-1)(\Delta(G) - 1 + 6 + 6 + 1)\}\\
&\leq 3(2p-1) + (2q-1)(M+12)\\
&= (2q-1)M + 6p + 24q - 15 = n_3.
\end{aligned}$$

If one of (C4), (C5), and (C6) holds, then erase $\phi(v)$ and $\phi(u)$ for any of its adjacent 2-vertices u. Afterwards, first label v, then label these adjacent 2-vertices, and finally label v_1.

It is easy to verify that, for cases (C4) to (C6),

$$\begin{aligned}
\sigma(u) &\leq 2(2p-1) + (2q-1)(M+5)\\
&= (2q-1)M + 4p + 10q - 7 < n_3
\end{aligned}$$

$$\text{if} \quad d(u)=2,$$
$$\sigma(v) \le \max\{2(2p-1)+(2q-1)(\Delta(G)+8),$$
$$(2p-1)+(2q-1)(\Delta(G)+4), 7(2q-1)\} \le n_3.$$

Thus ϕ can be extended to an $L(p,q)$-labeling of G with the label set L_3. □

4. Concluding remarks. Since both $\chi(G^2)$ and $\lambda(G;2,1)$ are greater than or equal to $\Delta(G)+1$, it follows from Corollary 2 that each of them is equal to the maximum degree plus a constant if G is a planar graph of girth at least 5. Let $c'(g)$ denote the smallest integer k' such that all planar graphs G of girth at least g satisfy $\chi(G^2) \le \Delta(G)+k'$. Let $c''(g)$ be defined similarly with respect to $\lambda(G;2,1)$. Both $c'(g)$ and $c''(g)$ are well defined when $g \ge 5$. A 5-cycle C_5 satisfies $\chi(C_5^2) = 5 = \Delta(C_5)+3$ and $\lambda(C_5;2,1) = 4 = \Delta(C_5)+2$; a 7-cycle C_7 satisfies $\chi(C_7^2) = \lambda(C_7;2,1) = 4 = \Delta(C_7)+2$. Let H denote the graph obtained by inserting a new vertex into each of the edges of a complete graph on four vertices. It is easy to verify that $\chi(H^2) = 5 = \Delta(H)+2$ and $\lambda(H;2,1) = 6 = \Delta(H)+3$. These examples together with Corollary 2 show that

(i) $3 \le c'(5) \le 16$ and $2 \le c''(5) \le 21$;
(ii) $2 \le c'(6) \le 10$ and $3 \le c''(6) \le 15$;
(iii) $2 \le c'(7) \le 5$ and $2 \le c''(7) \le 8$.

Given $g \ge 5$, determining the precise values of $c'(g)$ and $c''(g)$ seems to be an interesting problem.

We remark that neither $c'(g)$ nor $c''(g)$ is well defined when $g \le 4$. Infinitely many counterexamples have been constructed in [9].

In conclusion, we would like to propose the following.

CONJECTURE 3. *For any integer $g \ge 5$, there exists a sufficiently large integer $M(g)$ such that if G is a planar graph with girth g and $\Delta(G) \ge M(g)$, then $\chi(G^2) = \lambda(G;2,1) = \Delta(G)+1$.*

Acknowledgment. The authors are very grateful to an anonymous referee for giving them detailed comments and directing their attention to [12].

REFERENCES

[1] O. V. Borodin, H. J. Broersma, A. Glebov, and J. van den Heuvel, *Stars and Bunches in Planar Graphs. Part* II: *General Planar Graphs and Colourings*, CDAM research report, London School of Economics, London, 2002, translated and adapted from Diskretn. Anal. Issled. Oper. Ser. 1, 8 (2001), pp. 9–33 (in Russian).

[2] G. J. Chang and D. Kuo, *The $L(2,1)$-labeling problem on graphs*, SIAM J. Discrete Math., 9 (1996), pp. 309–316.

[3] G. J. Chang, W.-T. Ke, D. Kuo, D. D.-F. Liu, and R. K. Yeh, *On $L(d,1)$-labeling of graphs*, Discrete Math., 220 (2000), pp. 57–66.

[4] J. P. Georges and D. W. Mauro, *Generalized vertex labeling with a condition at distance two*, Congr. Numer., 109 (1995), pp. 141–159.

[5] J. P. Georges, D. W. Mauro, and M. A. Whittlesey, *Relating path coverings to vertex labellings with a condition at distance two*, Discrete Math., 135 (1994), pp. 103–111.

[6] J. R. Griggs and R. K. Yeh, *Labelling graphs with a condition at distance* 2, SIAM J. Discrete Math., 5 (1992), pp. 586–595.

[7] J. van den Heuvel and S. McGuinness, *Coloring the square of a planar graph*, J. Graph Theory, 42 (2003), pp. 110–124.

[8] P. K. Jha, A. Narayanan, P. Sood, K. Sundaram, and V. Sunder, *On $L(2,1)$-labeling of the Cartesian product of a cycle and a path*, Ars Combin., 55 (2000), pp. 81–89.

[9] K.-W. Lih, W.-F. Wang, and X. Zhu, *Coloring the square of a K_4-minor free graph*, Discrete Math., 269 (2003), pp. 303–309.

[10] K.-W. Lih and W.-F. Wang, *Coloring the Square of an Outerplanar Graph*, preprint, Academia Sinica, Taipei, Taiwan, 2002.

[11] D. Sakai, *Labeling chordal graphs: Distance two condition*, SIAM J. Discrete Math., 7 (1994), pp. 133–140.

[12] C. Thomassen, *Applications of Tutte Cycles*, Technical report, Technical University of Denmark, Copenhagen, 2001.

[13] G. Wegner, *Graphs with Given Diameter and a Coloring Problem*, preprint, University of Dortmund, Dortmund, Germany, 1977.

[14] M. A. Whittlesey, J. P. Georges, and D. W. Mauro, *On the λ-number of Q_n and related graphs*, SIAM J. Discrete Math., 8 (1995), pp. 499–506.

SIAM J. Discrete Math.
Vol. 17, No. 2, pp. 276–297

EQUALITIES AMONG CAPACITIES OF (d, k)-CONSTRAINED SYSTEMS*

NAVIN KASHYAP† AND PAUL H. SIEGEL†

Abstract. In this paper, we consider the problem of determining when the capacities of distinct (d, k)-constrained systems can be equal. A (d, k)-constrained system consists of binary sequences which have at least d zeros and at most k zeros between any two successive ones. If we let $C(d, k)$ denote the capacity of a (d, k)-constrained system, then it is known that $C(d, 2d) = C(d + 1, 3d + 1)$ and $C(d, 2d + 1) = C(d + 1, \infty)$. Repeated application of these two identities also yields the chain of equalities $C(1, 2) = C(2, 4) = C(3, 7) = C(4, \infty)$. We show that these are the only equalities possible among the capacities of (d, k)-constrained systems. In the process, we also provide useful factorizations of the characteristic polynomials for these constraints.

Key words. Shannon capacity, constrained systems, (d, k)-constraints, polynomial factorization

AMS subject classifications. 94A55, 11R09

DOI. 10.1137/S0895480102413710

1. Introduction.

Given nonnegative integers d, k, with $d < k$, we say that a binary sequence is (d, k)-constrained if every run of zeros has length at most k and any two successive ones are separated by a run of zeros of length at least d. A (d, k)-*constrained system* is defined to be the set of all finite-length (d, k)-constrained binary sequences. The above definition can be extended to the case $k = \infty$ by not imposing an upper bound on the lengths of zero-runs. In other words, a binary sequence is said to be (d, ∞)-constrained if any two successive ones are separated by at least d zeros, and a (d, ∞)-*constrained system* is defined to be the set of all finite-length (d, ∞)-constrained binary sequences. From now on, when we refer to (d, k)-constrained systems, we shall also allow k to be ∞.

Let $\mathcal{S}(d, k)$ be a (d, k)-constrained system, and let $q_{d,k}(n)$ be the number of length-n sequences in $\mathcal{S}(d, k)$. The *Shannon capacity*, or simply *capacity*, of $\mathcal{S}(d, k)$ is defined as

$$C(d, k) = \lim_{n\to\infty} \frac{1}{n} \log_2 q_{d,k}(n). \tag{1}$$

It is well known (see, e.g., [2]) that $C(d, k) = \log_2 \rho_{d,k}$, where $\rho_{d,k}$ is the unique largest-magnitude root of a certain polynomial, $\chi_{d,k}(z)$, called the *characteristic polynomial* of the constraint. When k is finite, $\chi_{d,k}(z)$ takes the form

$$\chi_{d,k}(z) = z^{k+1} - \sum_{j=0}^{k-d} z^j, \tag{2}$$

and when $k = \infty$,

$$\chi_{d,\infty}(z) = z^{d+1} - z^d - 1. \tag{3}$$

$\rho_{d,k}$ is always real and lies in the interval $(1, 2]$ so that $0 < C(d, k) \leq 1$. In fact, $C(d, k) = 1$ if and only if $(d, k) = (0, \infty)$.

*Received by the editors September 3, 2002; accepted for publication (in revised form) June 22, 2003; published electronically November 14, 2003.
http://www.siam.org/journals/sidma/17-2/41371.html

†Department of Electrical and Computer Engineering, University of California, San Diego, CA 92093-0407 (nkashyap@ece.ucsd.edu, psiegel@ece.ucsd.edu).

Interest in constrained systems and their capacities dates back to the work of Shannon [8]. In the mathematical literature, constrained systems are the subject of study of symbolic dynamics (cf. [3]), where the capacity of a constrained system is referred to as its entropy. (d,k)-constrained systems in particular have applications in magnetic and optical recording systems [5].

It is easily verified that certain pairs of (d,k)-constrained systems have the same capacity. For example, we have the identities

$$C(d,2d) = C(d+1,3d+1), \tag{4}$$

$$C(d,2d+1) = C(d+1,\infty) \tag{5}$$

true for all $d \geq 0$. The first equality is a consequence of the fact that $\chi_{d+1,3d+1}(z) = (z^{d+1}+1)\,\chi_{d,2d}(z)$, since all the roots of $z^{d+1}+1$ lie on the unit circle so that $\rho_{d,2d} = \rho_{d+1,3d+1}$. Similarly, the factorization $\chi_{d,2d+1}(z) = \chi_{d+1,\infty}(z)\sum_{i=0}^{d} z^i$ yields (5), since $\sum_{i=0}^{d} z^i = (z^{d+1}-1)/(z-1)$ has all its roots on the unit circle as well.

Repeatedly applying the two identities above also yields the chain of equalities

$$C(1,2) = C(2,4) = C(3,7) = C(4,\infty). \tag{6}$$

It is the aim of this paper to show that (4), (5), and (6) capture all the equalities possible among the capacities of (d,k)-constrained systems. More precisely, we shall prove the following theorem.

Theorem 1. *If $C(d,k) = C(\hat{d},\hat{k})$ for $(d,k) \neq (\hat{d},\hat{k})$, then one of the following holds:*

(i) $\{(d,k),(\hat{d},\hat{k})\} = \{(\ell,2\ell),\,(\ell+1,3\ell+1)\}$ *for some integer* $\ell \geq 0$,

(ii) $\{(d,k),(\hat{d},\hat{k})\} = \{(\ell,2\ell+1),\,(\ell+1,\infty)\}$ *for some integer* $\ell \geq 0$,

(iii) $(d,k),(\hat{d},\hat{k})$ *are among the pairs listed in* (6).

The key to our proof of this result is an explicit factorization we obtain for the characteristic polynomials of the (d,k)-constraints. We show that $\chi_{d,k}(z)$ can be factored as

$$\chi_{d,k}(z) = \Phi_{d,k}(z)\,\Psi_{d,k}(z),$$

where $\Phi_{d,k}(z), \Psi_{d,k}(z) \in \mathbb{Z}[z]$, $\Psi_{d,k}(z)$ is irreducible (over $\mathbb{Z}$), and $\Phi_{d,k}(z)$ either is 1 or has all its roots on the unit circle. We can, in fact, determine an explicit form for the polynomials $\Phi_{d,k}(z)$, from which we can deduce an expression for $\Psi_{d,k}(z)$ for certain (d,k) pairs. An immediate consequence of this result is that $C(d,k) = C(\hat{d},\hat{k})$ if and only if $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$. Theorem 1 is then obtained by identifying all the cases where we can have $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$. This last step relies heavily on the explicit form we derive for the Φ and Ψ polynomials.

The rest of the paper is organized as follows. In section 2, we present the factorization of $\chi_{d,k}(z)$, which we use in section 3 to prove Theorem 1.

2. Factorization of $\chi_{d,k}(z)$.

We shall first consider the factorization of $\chi_{d,\infty}(z)$, as it follows directly from existing results. Throughout this paper, we shall be concerned only with polynomials with integer coefficients. Any such polynomial is called reducible if it can be factored over the integers, and irreducible otherwise.

If $F(z) \in \mathbb{Z}[z]$ is a polynomial of degree n, then $F^*(z) = z^n F(1/z)$ is called the *reciprocal polynomial* of $F(z)$. Thus, for example, if $F(z) = z^5 - 4z^4 + 6z^3 - 4z^2 - 1$, then $F^*(z) = 1 - 4z + 6z^2 - 4z^3 - z^5$ is its reciprocal polynomial.

Observe that $\chi^*_{d,\infty}(z) = 1 - z - z^{d+1}$, so that when d is odd, $-\chi^*_{d,\infty}(-z) = z^{d+1} - z - 1$, and when d is even, $\chi^*_{d,\infty}(-z) = z^{d+1} + z + 1$. The following result deals with the irreducibility of the polynomials $z^n - z - 1$ and $z^n + z + 1$.

THEOREM 2 (see [7, Theorem 1]). (i) $z^n - z - 1$ *is irreducible for all* n. (ii) *For* $n > 2$, $z^n + z + 1$ *is irreducible if and only if* $n \not\equiv 2 \pmod 3$. *If* $n \equiv 2 \pmod 3$, *then* $z^2 + z + 1$ *is a factor and the other factor is irreducible.*

Thus, by part (i) of the above theorem, for odd d, $-\chi^*_{d,\infty}(-z)$ is irreducible and hence so is $\chi_{d,\infty}(z)$. When d is even, it is either 0, 2, or 4 $\pmod 6$. In the first two cases, $d+1 \not\equiv 2 \pmod 3$, and so by part (ii) of the above result, $\chi^*_{d,\infty}(-z)$ is irreducible and therefore so is $\chi_{d,\infty}(z)$. When $d \equiv 4 \pmod 6$, we have $d + 1 \equiv 2 \pmod 3$, and applying part (ii) of the theorem again, we see that $\chi^*_{d,\infty}(-z) = (z^2 + z + 1)p(z)$ for some irreducible $p(z)$. Therefore, in this case, we have $\chi_{d,\infty}(z) = (z^2 - z + 1)\Psi_{d,\infty}(z)$, with $\Psi_{d,\infty}(z) = p^*(-z)$ being irreducible. In fact, one can easily verify by means of an inductive argument that when $d \equiv 4 \pmod 6$, then

$$\Psi_{d,\infty}(z) = z^3 - z - 1 + \sum_{l=2}^{(d+2)/6} (z^{6l-3} - z^{6l-5} - z^{6l-6} + z^{6l-8}). \tag{7}$$

We summarize these results in the following theorem.

THEOREM 3. *For* $d \not\equiv 4 \pmod 6$, $\chi_{d,\infty}(z)$ *is irreducible. For* $d \equiv 4 \pmod 6$, $\chi_{d,\infty}(z) = (z^2 - z + 1)\Psi_{d,\infty}(z)$, *with* $\Psi_{d,\infty}(z)$ *irreducible and of the form given by* (7).

When k is finite, the factorization we obtain for $\chi_{d,k}(z)$ is based on a technique originally due to Ljunggren [4], which was further developed by Filaseta [1]. We briefly describe this technique here.

We define $F(z) \in \mathbb{Z}[z]$ to be *self-reciprocal* if $F(z) = \pm F^*(z)$. Note that $F(z)$ is self-reciprocal if and only if λ being a root of $F(z)$ implies that λ^{-1} is also a root. An example of a polynomial that is self-reciprocal is $z^5 - 10z^3 + 10z^2 - 1$.

Now, any $F(z) \in \mathbb{Z}[z]$ can always be written as $F(z) = \Phi(z)\Psi(z)$, where $\Phi(z)$ is the product of all the irreducible self-reciprocal factors of $F(z)$ that have positive leading coefficients. If $F(z)$ has no irreducible self-reciprocal factors, then we take $\Phi(z) = 1$ and $\Psi(z) = F(z)$. We call $\Phi(z)$ the *reciprocal part* of $F(z)$, while $\Psi(z)$ is called the *nonreciprocal part* of $F(z)$. It is worth pointing out that this definition does not preclude $\Psi(z)$ from being self-reciprocal itself. For example, $F(z) = z^6 + z^5 + z^4 + 3z^3 + z^2 + z + 1 = (z^3 + z^2 + 1)(z^3 + z + 1)$, and both the factors are irreducible but not self-reciprocal. Thus, the nonreciprocal part of $F(z)$ is $F(z)$ itself, which is a self-reciprocal polynomial. On the other hand, the reciprocal part of any polynomial is always self-reciprocal.

Note that if we take $F(z) = \chi_{d,\infty}(z)$, then Theorem 3 shows that the reciprocal part of $F(z)$ is 1 when $d \not\equiv 4 \pmod 6$ and is $z^2 - z + 1$ when $d \equiv 4 \pmod 6$. Thus, the nonreciprocal part of $F(z)$ is $F(z)$ itself in the former case and is $\Psi_{d,\infty}(z)$ as given by (7) in the latter case. Observe that in either case, the nonreciprocal part of $F(z)$ is irreducible.

The following result [1, Lemma 1] tells us precisely when the nonreciprocal part of a polynomial is reducible.

LEMMA 4 (Ljunggren–Filaseta lemma). *The nonreciprocal part of* $F(z) \in \mathbb{Z}[z]$ *is reducible if and only if there exists* $G(z)$ *different from* $\pm F(z)$ *and* $\pm F^*(z)$ *such that* $G(z)G^*(z) = F(z)F^*(z)$.

The "only if" part of this lemma is sufficient for our purposes. To verify this part, note that if the nonreciprocal part, $\Psi(z)$, is reducible, then $\Psi(z) = A(z)B(z)$ for some non-self-reciprocal polynomials $A(z)$ and $B(z)$. Setting $G(z) = A(z)B^*(z)\Phi(z)$, where $\Phi(z)$ is the reciprocal part of $F(z)$, we see that $G(z)$ has the properties stated in the lemma.

We shall use the Ljunggren–Filaseta lemma to prove the irreducibility of the nonreciprocal part of $\chi_{d,k}(z)$ $(k < \infty)$. Once this is done, we shall study the reciprocal part of the polynomial. Recall that $\chi_{d,k}(z)$ is a polynomial of the form $f(z) = z^n - z^m - z^{m-1} - \cdots - z - 1$ for some $n > m > 0$. It is well known (see, e.g., [9]) that when $n = m+1$, the polynomial $f(z)$ is itself irreducible. So, we need only consider the case when $n \geq m+2$. We shall show that if $g(z) \in \mathbb{Z}[z]$ is a polynomial such that $g(z)g^*(z) = f(z)f^*(z)$, then $g(z) = \pm f(z)$ or $\pm f^*(z)$. The "only if" part of the Ljunggren–Filaseta lemma then shows that the nonreciprocal part of $f(z)$ is irreducible.

So, let $g(z) = \sum_{i=0}^{n} g_i z^i$ be a polynomial in $\mathbb{Z}[z]$ such that $g(z)g^*(z) = f(z)f^*(z)$. Note that $g(z)$ must itself be a polynomial of degree n. Without loss of generality, we may assume that $g_n > 0$ (else, replace $g(z)$ by $-g(z)$).

Lemma 5. *The coefficients g_i of $g(z)$ must satisfy the following equations:*

$$g_n = 1\,, \qquad g_0 = -1, \tag{8}$$

$$g_1 - g_{n-1} = -1, \tag{9}$$

$$\sum_{i=1}^{n-2} g_i g_{i+1} = m - 1, \tag{10}$$

$$\sum_{i=1}^{n-1} g_i^2 = m. \tag{11}$$

Proof. Let $f(z) = \sum_{i=0}^{n} f_i z^i$ so that $f_n = 1$, $f_i = 0$ for $m+1 \leq i \leq n-1$, and $f_i = -1$ for $0 \leq i \leq m$.

Equating the constant coefficients of $f(z)f^*(z)$ and $g(z)g^*(z)$, we see that $g_0 g_n = -1$. Since $g_0, g_n \in \mathbb{Z}$ and $g_n > 0$, we must have $g_n = 1, g_0 = -1$.

(9) is obtained by equating the coefficients of z in $f(z)f^*(z)$ and $g(z)g^*(z)$. The coefficient of z in $g(z)g^*(z)$ is $g_0 g_{n-1} + g_1 g_n = g_1 - g_{n-1}$. Now, note that since $n \geq m+2$, we have $f_{n-1} = 0$. Hence, the coefficient of z in $f(z)f^*(z)$ is $f_0 f_{n-1} + f_1 f_n = -1$.

To get (10), we equate the coefficients of z^{n-1}. In $g(z)g^*(z)$, this coefficient is $\sum_{i=0}^{n-1} g_i g_{i+1}$, while in $f(z)f^*(z)$, it is $\sum_{i=0}^{n-1} f_i f_{i+1} = \sum_{i=0}^{m-1} f_i f_{i+1}$, since $f_{i+1} = 0$ for $m \leq i \leq n-2$, and $f_i = 0$ for $i = n-1$. But in the range $0 \leq i \leq m-1$, $f_i = f_{i+1} = -1$, which shows that $\sum_{i=0}^{m-1} f_i f_{i+1} = m$. Thus, we have $\sum_{i=0}^{n-1} g_i g_{i+1} = m$, which reduces to (10) upon using (8) and (9).

Finally, the coefficient of z^n in $g(z)g^*(z)$ is $\sum_{i=0}^{n} g_i^2$, and correspondingly, in $f(z)f^*(z)$ is $\sum_{i=0}^{n} f_i^2 = m+2$. Hence, $\sum_{i=0}^{n} g_i^2 = m+2$, and since $g_0^2 = g_n^2 = 1$, we see that $\sum_{i=1}^{n-1} g_i^2 = m$, which proves (11). □

We use this lemma to prove the following proposition.

Proposition 6. *The nonreciprocal part of $f(z) = z^n - z^m - z^{m-1} - \cdots - z - 1$, $n > m > 0$, is irreducible.*

Proof. As noted above, we need only prove the result for $n \geq m+2$. Lemma 5 (which applies for $n \geq m+2$) shows that any $g(z) = \sum_{i=0}^{n} g_i z^i$ such that $g(z)g^*(z) =$

$f(z)f^*(z)$ and $g_n > 0$ must satisfy (8)–(11). Now, observe that

$$\begin{aligned}\sum_{i=1}^{n-2}(g_i - g_{i+1})^2 &= \sum_{i=1}^{n-2} g_i^2 + \sum_{i=1}^{n-2} g_{i+1}^2 - 2\sum_{i=1}^{n-2} g_i g_{i+1} \\ &= 2\sum_{i=1}^{n-1} g_i^2 - g_1^2 - g_{n-1}^2 - 2\sum_{i=1}^{n-2} g_i g_{i+1} \\ &= 2m - g_1^2 - g_{n-1}^2 - 2(m-1)\end{aligned}$$

with the last equality using (10) and (11). Thus, we see that

$$g_1^2 + g_{n-1}^2 + \sum_{i=1}^{n-2}(g_i - g_{i+1})^2 = 2. \tag{12}$$

Since all the g_i's are integers, this equation is satisfied if and only if exactly $n-2$ of the quantities g_1, g_{n-1}, $g_i - g_{i+1}$ ($i = 1, 2, \ldots, n-2$) are 0, and the remaining two nonzero quantities take values from the set $\{-1, 1\}$. In particular, $g_1 \in \{-1, 0, 1\}$. We consider each of the three choices for g_1 in turn.

If $g_1 = -1$, then (9) shows that $g_{n-1} = 0$. Hence, there exists a $k \in \{1, 2, \ldots, n-2\}$ such that $g_k - g_{k+1} = \pm 1$ and $g_i - g_{i+1} = 0$ for $i = 1, 2, \ldots, n-2$, $i \neq k$. Now, if $g_k - g_{k+1} = 1$, then we must have $g_i = -1$ for $1 \leq i \leq k$, and $g_i = -2$ for $k+1 \leq i \leq n-1$, which contradicts $g_{n-1} = 0$. Hence, $g_k - g_{k+1}$ must be -1, in which case $g_i = -1$ for $1 \leq i \leq k$, and $g_i = 0$ for $k+1 \leq i \leq n-1$. Using (11), we see that $k = m$, which forces $g(z)$ to be $z^n - z^m - z^{m-1} - \cdots - z - 1 = f(z)$.

If $g_1 = 0$, then (9) yields $g_{n-1} = 1$. As above, we must have $g_k - g_{k+1} = \pm 1$ for some $k \in \{1, 2, \ldots, n-2\}$, and $g_i - g_{i+1} = 0$ for $i = 1, 2, \ldots, n-2$, $i \neq k$. This time, choosing $g_k - g_{k+1}$ to be 1 leads to $g_{n-1} = -1$, which contradicts $g_{n-1} = 1$. Thus, $g_k - g_{k+1} = -1$, so that $g_i = 0$ for $1 \leq i \leq k$, and $g_i = 1$ for $k+1 \leq i \leq n-1$. From (11), we now get $k+1 = n-m$. Hence, $g(z)$ must be $z^n + z^{n-1} + \cdots + z^{n-m} - 1 = -f^*(z)$.

If $g_1 = 1$, then (9) implies that $g_{n-1} = 2$, which means that (12) cannot be satisfied. So, g_1 cannot be 1.

Thus, we have shown that if $g(z)$ is such that $g(z)g^*(z) = f(z)f^*(z)$ and $g_n > 0$, then $g(z) = f(z)$ or $g(z) = -f^*(z)$. For any $g(z)$ with $g_n < 0$, we can apply the above reasoning to $-g(z)$. This proves that if $g(z) \in \mathbb{Z}[z]$ is such that $g(z)g^*(z) = f(z)f^*(z)$, then $g(z) = \pm f(z)$ or $\pm f^*(z)$. The proposition now follows from the Ljunggren–Filaseta lemma. □

Having shown the irreducibility of the nonreciprocal part of $f(z) = z^n - z^m - z^{m-1} - \cdots - z - 1$, we move on to analyzing the reciprocal part, $\phi(z)$, of $f(z)$. Our first goal is to show that all the roots of $\phi(z)$ are in fact certain roots of unity, which will help us in determining the exact form of $\phi(z)$.

Lemma 7. *If λ is a root of $\phi(z)$, then λ is a root of either $\sum_{i=0}^{m-1} z^i$ or $\sum_{i=0}^{m+1} z^i$. In other words, λ is either an mth or an $(m+2)$nd root of unity, distinct from* 1.

Proof. Let λ be a root of $\phi(z)$. Note that $\lambda \neq 0$ because 0 cannot be a root of $f(z)$, as $f(0) = -1$. Since $\phi(z)$ is a self-reciprocal polynomial, λ^{-1} is also a root of $\phi(z)$. Since $\phi(z)$ is a factor of $f(z)$, we have $f(\lambda) = f(\lambda^{-1}) = 0$. This implies that

$$\lambda^n - \lambda^m - \lambda^{m-1} - \cdots - \lambda - 1 = 0, \tag{13}$$

$$\lambda^n + \lambda^{n-1} + \cdots + \lambda^{n-m+1} + \lambda^{n-m} - 1 = 0. \tag{14}$$

Equating the left-hand sides of these two equations, cancelling out the common terms, and rearranging, we obtain

$$(\lambda^{n-1}+\lambda^{n-2}+\cdots+\lambda^{n-m})+(\lambda^m+\lambda^{m-1}+\cdots+\lambda)=0.$$

Dividing through by $\lambda \neq 0$, we see that the above equation simplifies to

$$(\lambda^{n-m-1}+1)(\lambda^{m-1}+\lambda^{m-2}+\cdots+1)=0.$$

Hence λ is a root of either $z^{n-m-1}+1$ or $\sum_{i=0}^{m-1} z^i$. However, if λ is a root of $z^{n-m-1}+1$, then $\lambda^{n-m-1}=-1$. Now, note that (14) can be rewritten as $\lambda^{n-m-1}(\lambda^{m+1}+\lambda^m+\cdots+\lambda)-1=0$, which reduces to $-\lambda^{m+1}-\lambda^m-\cdots-\lambda-1=0$, since $\lambda^{n-m-1}=-1$. Hence if λ is a root of $z^{n-m-1}+1$, then it is also a root of $\sum_{i=0}^{m+1} z^i$, which proves the lemma. □

We can actually say something more about the roots of $\phi(z)$, as we shall see in the next few lemmas.

LEMMA 8. *If λ is a root of $\phi(z)$ that is also a root of $\sum_{i=0}^{m-1} z^i$, then λ is in fact a root of $\sum_{i=0}^{q-1} z^i$, where $q=\gcd(m,n)$.*

Proof. Suppose that λ is as in the hypothesis of the lemma. Since $\phi(\lambda)=0$, we also have $f(\lambda)=0$, which means that

$$\lambda^n-\sum_{i=0}^{m}\lambda^i=0. \tag{15}$$

But since λ is a root of $\sum_{i=0}^{m-1} z^i$, we have $\sum_{i=0}^{m-1}\lambda^i=0$ and, moreover, $\lambda^m=1$. Hence (15) reduces to $\lambda^n=1$. Hence λ is also an nth root of unity distinct from 1, i.e., λ is a root of $\sum_{i=0}^{n-1} z^i$. Therefore, λ is a root of $\gcd(\sum_{i=0}^{m-1} z^i, \sum_{i=0}^{n-1} z^i)=\sum_{i=0}^{q-1} z^i$, where $q=\gcd(m,n)$. □

When λ is an $(m+2)$nd root of unity, things get a little more complicated.

LEMMA 9. *If λ is a root of $\phi(z)$ that is also a root of $\sum_{i=0}^{m+1} z^i$, then* (i) *m is even,* (ii) *λ is a root of z^r+1, where $r=\gcd(\frac{m}{2}+1,n+1)$, and* (iii) *$(n+1)/r$ is odd.*

Proof. Let λ be as in the hypothesis of the lemma. Again, the fact that $f(\lambda)=0$ leads to (15). This time, since $\lambda\neq 1$ is an $(m+2)$nd root of unity, we have $\sum_{i=0}^{m+1}\lambda^i=0$, which implies that $-\sum_{i=0}^{m}\lambda^i=\lambda^{m+1}=1/\lambda$, using $\lambda^{m+2}=1$. Therefore, (15) reduces to $\lambda^n+1/\lambda=0$ or, equivalently, $\lambda^{n+1}=-1$.

Now, since λ is a root of $\sum_{i=0}^{m+1} z^i$, it is of the form $\lambda=e^{2\pi i\frac{k}{m+2}}$ for some $k\in\{1,2,\ldots,m+1\}$. Therefore, $-1=\lambda^{n+1}=e^{2\pi i\frac{k}{m+2}(n+1)}$. Hence $\frac{2k}{m+2}(n+1)=2j+1$ for some integer j, which upon rearrangement becomes

$$(2k)(n+1)=(2j+1)(m+2). \tag{16}$$

Since the left-hand side (LHS) of the above equation is even, so is the right-hand side (RHS). This means that m must be even, since $2j+1$ is odd. This proves (i).

Rearranging (16), we get $k\frac{n+1}{m/2+1}=2j+1$. Defining r to be $\gcd(\frac{m}{2}+1,n+1)$, we let $m'=(\frac{m}{2}+1)/r$ and $n'=(n+1)/r$. Thus, m',n' are integers such that $\gcd(m',n')=1$, and $\frac{n+1}{m/2+1}=\frac{n'}{m'}$. Therefore, we have

$$k\frac{n'}{m'}=2j+1. \tag{17}$$

Since $\gcd(m', n') = 1$, the fact that $k\frac{n'}{m'}$ is an integer implies that $m'|k$. Writing $k = lm'$ and plugging into (17), we get $ln' = 2j+1$. Therefore, $n'|(2j+1)$, which shows that n' is odd, thus proving (iii). Note that as $l|(2j+1)$, l is also odd.

Finally, $\lambda = e^{2\pi i \frac{k}{m+2}} = e^{\pi i \frac{k}{m/2+1}} = e^{\pi i \frac{lm'}{rm'}} = e^{\pi i \frac{l}{r}}$. Since l is odd, $\lambda^r = -1$, which shows that λ is a root of $z^r + 1$, thus completing the proof of the lemma. □

Now, from Lemmas 7, 8, and 9, we see that every root of $\phi(z)$ is also a root of $(\sum_{i=0}^{q-1} z^i)(z^r+1)$. In fact, for odd m, Lemma 9(i) shows that no root of $\phi(z)$ can be a root of z^r+1, so that every root of $\phi(z)$ is actually a root of $\sum_{i=0}^{q-1} z^i$. Now, if we can show that $\phi(z)$ has no repeated roots, it immediately follows that $\phi(z)$ is a factor of $\sum_{i=0}^{q-1} z^i$ for odd m, and of $(\sum_{i=0}^{q-1} z^i)(z^r+1)$ for even m. We proceed to show this next.

LEMMA 10. *$\phi(z)$ has no repeated roots.*

Proof. Suppose that λ is a repeated root of $\phi(z)$. Note that $|\lambda| = 1$ since any root of $\phi(z)$ is some root of unity. Define $g(z) = (z-1)f(z) = z^{n+1} - z^n - z^{m+1} + 1$. If λ is a repeated root of $\phi(z)$, then it must be a repeated root of $g(z)$ as well. Hence $g(\lambda) = g'(\lambda) = 0$, which implies that

$$\lambda^{n+1} - \lambda^n - \lambda^{m+1} + 1 = 0, \tag{18}$$

$$(n+1)\lambda^n - n\lambda^{n-1} - (m+1)\lambda^m = 0. \tag{19}$$

Multiplying (18) by $(n+1)$ and subtracting the result from λ times (19), we get

$$\lambda^n + (n-m)\lambda^{m+1} = n+1. \tag{20}$$

However, this leads to a contradiction because

$$n+1 = |\lambda^n + (n-m)\lambda^{m+1}| \le |\lambda|^n + (n-m)|\lambda|^{m+1} = 1 + n - m \le n,$$

with the last inequality arising from the fact that $m > 0$. This contradiction proves the lemma. □

As observed prior to the statement of Lemma 10, we can now conclude that $\phi(z)$ is a factor of $\sum_{i=0}^{q-1} z^i$ for odd m and of $(\sum_{i=0}^{q-1} z^i)(z^r+1)$ for even m.

In fact, for odd m, we can show that $\phi(z) = \sum_{i=0}^{q-1} z^i$. Since we already know that $\phi(z)|(\sum_{i=0}^{q-1} z^i)$ in this case, we need only to show that $(\sum_{i=0}^{q-1} z^i)|\phi(z)$. It actually suffices to show that $(\sum_{i=0}^{q-1} z^i)|f(z)$. This is because any factor, irreducible or otherwise, of $\sum_{i=0}^{q-1} z^i$ is always self-reciprocal (recall that $\phi(z)$ is the product of all irreducible self-reciprocal factors of $f(z)$): if $\pi(z)$ is a factor of $\sum_{i=0}^{q-1} z^i$ and λ is a root of $\pi(z)$, then so is its complex conjugate, $\overline{\lambda} = \lambda^{-1}$.

So, to show that $(\sum_{i=0}^{q-1} z^i)|f(z)$, we write $n = n'q$, $m = m'q$, so that

$$\begin{aligned} f(z) &= z^{n'q} - \sum_{i=0}^{m'q} z^i \\ &= z^{n'q} - z^{m'q} - \sum_{i=0}^{m'q-1} z^i \\ &= z^{m'q}(z^{(n'-m')q} - 1) - \left(\sum_{i=0}^{q-1} z^i\right)\left(\sum_{l=0}^{m'-1} z^{lq}\right) \end{aligned}$$

$$
\begin{aligned}
&= z^{m'q}(z^q-1)\left(\sum_{l=0}^{n'-m'-1} z^{lq}\right) - \left(\sum_{i=0}^{q-1} z^i\right)\left(\sum_{l=0}^{m'-1} z^{lq}\right) \\
&= z^{m'q}(z-1)\left(\sum_{i=0}^{q-1} z^i\right)\left(\sum_{l=0}^{n'-m'-1} z^{lq}\right) - \left(\sum_{i=0}^{q-1} z^i\right)\left(\sum_{l=0}^{m'-1} z^{lq}\right) \\
&= \left(\sum_{i=0}^{q-1} z^i\right)\left(z^{m'q}(z-1)\sum_{l=0}^{n'-m'-1} z^{lq} - \sum_{l=0}^{m'-1} z^{lq}\right) \\
&= \left(\sum_{i=0}^{q-1} z^i\right)\left(\sum_{l=m'}^{n'-1} z^{lq+1} - \sum_{l=0}^{n'-1} z^{lq}\right). \qquad (21)
\end{aligned}
$$

Thus, we have proved that $(\sum_{i=0}^{q-1} z^i)|f(z)$, which implies that $\phi(z) = \sum_{i=0}^{q-1} z^i$. Note that the factorization in (21) is true for any m and n, not just for odd m. However, odd m ensures that $\sum_{i=0}^{q-1} z^i$ is the reciprocal part of $f(z)$ and the other factor is the nonreciprocal part.

The above argument, in conjunction with Proposition 6, proves the following theorem.

Theorem 11. *Let $f(z) = z^n - \sum_{i=0}^{m} z^i$, $n > m > 0$, m odd, and let $q = \gcd(m,n)$. Then, $f(z) = (\sum_{i=0}^{q-1} z^i)\psi(z)$, with $\psi(z) = \sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq}$ irreducible. In particular, for odd m, $f(z)$ is irreducible if and only if $\gcd(m,n) = 1$.*

We next tackle the case when m is even, which is a little less clean. The first observation to be made here is that when $(n+1)/r$ is also even, where $r = \gcd(\frac{m}{2}+1, n+1)$, then it follows from Lemma 9(iii) that $\phi(z)$ cannot share any roots with $\sum_{i=0}^{m+1} z^i$. So, it must share all its roots with $\sum_{i=0}^{q-1} z^i$, q being $\gcd(m,n)$ as above, implying that $\phi(z)| \sum_{i=0}^{q-1} z^i$. So, applying the argument given prior to the statement of Theorem 11, we see that in this case as well, we have $f(z) = (\sum_{i=0}^{q-1} z^i)\psi(z)$, with $\psi(z)$ irreducible and of the form stated in the theorem. This situation holds, for example, when n is odd and $4|m$, since then $\frac{m}{2}+1$ is odd and so is r because $r|(\frac{m}{2}+1)$, leading to the conclusion that $(n+1)/r$ is even.

So, we are left with the case when m is even, but $(n+1)/r$ is odd. This is dealt with in the following proposition.

Proposition 12. *When m is even and $(n+1)/r$ is odd, then $\phi(z)$ is the least common multiple (lcm) of $\sum_{i=0}^{q-1} z^i$ and z^r+1.*

Proof. From Lemmas 7, 8, and 9, we know that $\phi(z)$ is a factor of $\phi_1(z)\phi_2(z)$, where we have defined $\phi_1(z) = z^r + 1$ and $\phi_2(z) = \sum_{i=0}^{q-1} z^i$. In fact, as $\phi(z)$ has no repeated roots, it must be a factor of $\frac{\phi_1(z)\phi_2(z)}{\gcd(\phi_1(z),\phi_2(z))} = \operatorname{lcm}(\phi_1(z), \phi_2(z))$, since dividing by $\gcd(\phi_1(z), \phi_2(z))$ takes out some roots common to $\phi_1(z)$ and $\phi_2(z)$.

So, we need to show the converse, i.e., that $\operatorname{lcm}(\phi_1(z), \phi_2(z))$ is a factor of $\phi(z)$. Equivalently, we need to show that $\phi_1(z)|\phi(z)$ and $\phi_2(z)|\phi(z)$. Recalling that $\phi(z)$ is the product of all the irreducible self-reciprocal factors of $f(z)$, it suffices to show that $\phi_1(z)|f(z)$ and $\phi_2(z)|f(z)$. This is because any factor, irreducible or otherwise, of either $\phi_1(z)$ or $\phi_2(z)$ is self-reciprocal. Indeed, if $\pi(z)$ is a factor of either polynomial and λ is a root of $\pi(z)$, then so is its complex conjugate $\overline{\lambda}$. But as λ, being a root of $\phi_1(z)$ or $\phi_2(z)$, lies on the unit circle, we have $\overline{\lambda} = \lambda^{-1}$, implying that $\pi(z)$ is self-reciprocal.

We have already seen (see (21)) that $\phi_2(z)|f(z)$. To prove that $\phi_1(z)|f(z)$, we shall show that $f(\lambda)=0$ for any root λ of $\phi_1(z)$, which is sufficient because $\phi_1(z)$ has no repeated roots. Since $\lambda \notin \{0,1\}$, it is enough to show that $\lambda(\lambda-1)f(\lambda)=0$, i.e., $\lambda^{n+2}-\lambda^{n+1}-\lambda^{m+2}+\lambda=0$. Now, $\lambda^{n+1}=(\lambda^r)^{(n+1)/r}=(-1)^{(n+1)/r}=-1$ as $(n+1)/r$ is odd. Moreover, defining $m'=(\frac{m}{2}+1)/r$, we have $\lambda^{m+2}=(\lambda^r)^{2m'}=(-1)^{2m'}=1$. Hence $\lambda^{n+2}-\lambda^{n+1}-\lambda^{m+2}+\lambda=-\lambda-(-1)-1+\lambda=0$, as desired. □

The next lemma explicitly determines the lcm of $\sum_{i=0}^{q-1} z^i$ and z^r+1.

Lemma 13. *If q is even, then*

$$\operatorname{lcm}\left(\sum_{i=0}^{q-1} z^i, z^r+1\right)=\frac{z^r+1}{z+1}\sum_{i=0}^{q-1} z^i=\left(\sum_{i=0}^{r-1}(-z)^i\right)\left(\sum_{i=0}^{q-1} z^i\right).$$

Otherwise,

$$\operatorname{lcm}\left(\sum_{i=0}^{q-1} z^i, z^r+1\right)=(z^r+1)\sum_{i=0}^{q-1} z^i.$$

Proof. Let $\phi_1(z)=z^r+1$ and $\phi_2(z)=\sum_{i=0}^{q-1} z^i$. Since $\gcd(\phi_1,\phi_2)\cdot\operatorname{lcm}(\phi_1,\phi_2)=\phi_1(z)\phi_2(z)$, the lemma is proved once we show that $\gcd(\phi_1,\phi_2)$ is $z+1$ if q is even, and 1 otherwise.

We first show that if $\gcd(\phi_1,\phi_2)\neq 1$ then q is even and $\gcd(\phi_1,\phi_2)=z+1$. Suppose that $\pi(z)$ is a nontrivial factor of both $\phi(z)$ and $\phi_2(z)$, so that there exists a λ such that $\phi_1(\lambda)=\phi_2(\lambda)=0$. Such a λ must be of the form $\lambda=e^{2\pi i\frac{k}{q}}$ for some $k\in\{1,2,\ldots,q-1\}$ and must satisfy $\lambda^r=-1$. Hence $e^{2\pi i\frac{kr}{q}}=-1$, which means that $2k\frac{r}{q}$ must be an odd integer.

Now, as $q|n$ and $r|(n+1)$, $\gcd(q,r)=1$. So, for $2k\frac{r}{q}$ to be an integer, $2k$ must be a multiple of q. Let $2k=ql$ so that $2k\frac{r}{q}=lr$. Thus, lr is an odd integer, which shows that r and l are both odd. Furthermore, since $2k=ql$, the fact that l is odd implies that q is even. In fact, this also forces λ to be -1, because $\lambda=e^{2\pi i\frac{k}{q}}=e^{\pi i l}=-1$, since l is odd.

Thus, if $\pi(z)$ is a nontrivial factor of both $\phi_1(z)$ and $\phi_2(z)$, then $\lambda=-1$ is the only root that $\pi(z)$ can have. Since neither $\phi_1(z)$ nor $\phi_2(z)$ has repeated roots, -1 must be a simple root of $\pi(z)$, which shows that $\pi(z)=z+1$. We have thus shown that if $\gcd(\phi_1,\phi_2)$ is nontrivial, then q is even and $\gcd(\phi_1,\phi_2)=z+1$.

It remains to show only that if q is even, then $\gcd(\phi_1,\phi_2)=z+1$. Note that if $q=\gcd(m,n)$ is even, then so is n. Therefore, $n+1$ is odd, and since $r|(n+1)$, so is r. But, for even q and odd r, it is clear that $\phi_1(-1)=\phi_2(-1)=0$. Hence $(z+1)|\gcd(\phi_1,\phi_2)$, meaning that $\gcd(\phi_1,\phi_2)$ is nontrivial. But as we have already shown, this implies that $\gcd(\phi_1,\phi_2)=z+1$. □

We compile all the results proved above for the case when m is even in the following theorem.

Theorem 14. *Let $f(z)=z^n-\sum_{i=0}^{m} z^i$, $n>m>0$, m even, and let $q=\gcd(m,n)$, $r=\gcd(\frac{m}{2}+1,n+1)$, $n'=(n+1)/r$. Then, $f(z)=\phi(z)\psi(z)$, where $\psi(z)$ is irreducible and*

$$\phi(z)=\begin{cases}\sum_{i=0}^{q-1} z^i & \text{if } n' \text{ is even,}\\ \left(\sum_{i=0}^{r-1}(-z)^i\right)\left(\sum_{i=0}^{q-1} z^i\right) & \text{if } q \text{ is even,}\\ (z^r+1)\sum_{i=0}^{q-1} z^i & \text{otherwise.}\end{cases}$$

We would like to remark that when q is even, $n' = (n+1)/r$ is odd, so that the statement of the theorem is indeed consistent.

At this stage, it is worth pointing out that the results of Theorems 11 and 14 can be partially obtained from results in the existing literature, specifically [4] and [6]. Observe that, as noted in the proof of Lemma 10, we may define $g(z) = (z-1)f(z) = z^{n+1} - z^n - z^{m+1} + 1$. Now, Ljunggren [4] considered the factorization of polynomials of the form $q(x) = x^n \pm x^m \pm x^p \pm 1$ with $n > m > p > 0$ and claimed to show that all such polynomials can be factored as $q(x) = \phi(x)\psi(x)$, where $\phi(x)$ is self-reciprocal and has all its zeros on the unit circle and $\psi(x)$ is either 1 or a non-self-reciprocal irreducible polynomial. However, there was a minor error in Ljunggren's work, which was subsequently corrected by Mills [6]. Mills's work shows that Ljunggren's claim is in fact true for any polynomial $g(z)$ as above. Since $m+1 \geq 2$, $g(z)$ is not self-reciprocal and hence must have a nontrivial nonreciprocal part $\psi(z)$. Thus, these results show that $g(z)$, and hence $f(z)$, can be written as the product of a self-reciprocal polynomial having all its roots on the unit circle and a nontrivial, irreducible, non-self-reciprocal polynomial. Of course, these results do not go so far as to provide the specific forms of the reciprocal and nonreciprocal parts of $f(z)$ that we have derived above. So, in the interest of keeping our paper self-contained, we have chosen to include complete proofs of the aforementioned theorems.

3. Identifying equalities among (d,k) capacities. We shall use the factorization obtained in the previous section for the characteristic polynomials of (d,k) constraints to determine all possible equalities among the capacities of such constraints. We begin by showing that this problem is equivalent to the one of determining when the nonreciprocal parts of the characteristic polynomials of two such constraints can be equal. Throughout this section, we consider (d,k) pairs such that $0 < d < k \leq \infty$, and $\Phi_{d,k}(z)$ and $\Psi_{d,k}(z)$ will be used to denote the reciprocal and nonreciprocal parts, respectively, of the characteristic polynomial $\chi_{d,k}(z)$. Also, given polynomials $f(z), g(z)$, we shall use $f(z) = g(z)$ to denote that the two polynomials are identical.

THEOREM 15. *$C(d,k) = C(\hat{d},\hat{k})$ if and only if $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$.*

Proof. We shall show that $\rho_{d,k} = \rho_{\hat{d},\hat{k}}$ if and only if $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$, the ρ's being the largest roots of their respective characteristic polynomials.

Observe first that since the reciprocal parts of the characteristic polynomials have all their roots on the unit circle, and the ρ's are strictly greater than 1, the ρ's must be roots of the nonreciprocal parts. So, if $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$, then their largest roots must be identical, i.e., $\rho_{d,k} = \rho_{\hat{d},\hat{k}}$.

Conversely, suppose that $\rho_{d,k} = \rho_{\hat{d},\hat{k}}$. Since $\Psi_{d,k}(z)$ is irreducible and has $\rho_{d,k}$ as a root, it must be the minimal polynomial (over $\mathbb{Z}$) of $\rho_{d,k}$. Similarly, $\Psi_{\hat{d},\hat{k}}(z)$ is the minimal polynomial of $\rho_{\hat{d},\hat{k}}$. Hence by the uniqueness of the minimal polynomial of an algebraic integer, $\rho_{d,k} = \rho_{\hat{d},\hat{k}}$ implies that $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$. $\square$

With this theorem in hand, we can begin our investigation of equalities among the capacities of (d,k)-constrained systems. We shall first consider the case when at least one of the (d,k) constraints has $k = \infty$. Observe that since $\Psi_{d,\infty}(z)$ is either $\chi_{d,\infty}(z)$ itself or of the form given in (7), we can have $C(d,\infty) = C(\hat{d},\infty)$, or equivalently, $\Psi_{d,\infty}(z) = \Psi_{\hat{d},\infty}(z)$, if and only if $d = \hat{d}$. So, we need only concern ourselves with the situation when $C(d,\infty) = C(\hat{d},\hat{k})$ with $\hat{k}$ finite.

At this point, we shall find it convenient to introduce some definitions.

Definition 16. *A polynomial $f(z) = z^n - \sum_{i=0}^{m} z^i$, $n > m > 0$, is defined as being of*

- Type I *if its reciprocal part, $\phi(z)$, is of the form $\sum_{i=0}^{q-1} z^i$, with $q \geq 1$ odd;*
- Type II *if $\phi(z)$ is of the form $(z^r + 1)\sum_{i=0}^{q-1} z^i$, with $q \geq 1$ odd, $r \geq 1$; and*
- Type III *if $\phi(z)$ is of the form $(\sum_{i=0}^{r-1}(-z)^i)(\sum_{i=0}^{q-1} z^i)$, with $q \geq 2$ even and $r \geq 3$ odd.*

Theorems 11 and 14 show that any such $f(z)$ is always of Type I, II, or III, with $q = \gcd(m, n)$ and $r = \gcd(\frac{m}{2} + 1, n + 1)$. These theorems can be used to determine exactly when $f(z)$ is of a particular type. For example, $f(z)$ is of Type I precisely when one of the following three conditions holds: (i) m is odd, (ii) m and $(n+1)/r$ are even, and (iii) m and n are even and $r = 1$. Note that when $f(z)$ is of Type I, its nonreciprocal part, $\psi(z)$, is of the form $\sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq}$, as shown by (21).

The following simple fact about $f(z)$'s of Type II or III will be used often.

Lemma 17. *Let m be even and let $f(z)$ be of Type* II *or* III. *If $q = \gcd(m, n)$ and $r = \gcd(\frac{m}{2} + 1, n + 1)$, then $q \neq r$.*

Proof. If $q = r$, then $f(z)$ cannot be of Type III, since the definition requires q to be even and r to be odd. So, suppose that $f(z)$ is of Type II, with $q = r$. Note that since $q|n$ and $r|(n + 1)$, we must have $\gcd(q, r) = 1$, and hence $q = r = 1$. As $z^r + 1 = z + 1$ is a factor of $f(z)$, we must have $f(-1) = 0$. Now, it is easily verified that since $f(z)$ has the form $z^n - \sum_{i=0}^{m} z^i$, $f(-1)$ can be 0 only if m and n are both even. So, $q = \gcd(m, n)$ is even, which is impossible since $q = 1$. □

We will also find the following set of definitions to be useful.

Definition 18. *Given a polynomial $g(z) = \sum_{k=0}^{n} c_k z^k$, we define*

- *$\epsilon_i(g)$, $i \geq 1$, to be the ith smallest $k > 0$ such that $c_k \neq 0$;*
- *$\xi_i(g)$, $i \geq 1$, to be the ith largest $k > 0$ such that $c_k \neq 0$.*

Thus, for example, with $g(z) = z^6 - z^3 - z^2 - z - 1$, we have $\epsilon_i(g) = i$ for $i = 1, 2, 3$, $\epsilon_4(g) = 6$, $\xi_1(g) = 6$, and $\xi_i(g) = 5 - i$ for $i = 2, 3, 4$. Note that if $g(z), h(z)$ are polynomials such that $g(z) = h(z)$, then $\epsilon_i(g) = \epsilon_i(h)$ and $\xi_i(g) = \xi_i(h)$ for all $i \geq 1$.

We tackle the equality $C(d, \infty) = C(\hat{d}, \hat{k})$ through a series of lemmas, each of which considers a special case in which $\chi_{d,\infty}(z)$ is either irreducible ($d \not\equiv 4 \pmod 6$) or reducible ($d \equiv 4 \pmod 6$), and $\chi_{\hat{d},\hat{k}}(z)$ is of one of the three types defined above.

Lemma 19. *Let $d \not\equiv 4 \pmod 6$ and $\hat{d}, \hat{k}$ be such that $\chi_{\hat{d},\hat{k}}(z)$ is of Type* I. *Then, $C(d, \infty) = C(\hat{d}, \hat{k})$ only if $(\hat{d}, \hat{k}) = (d - 1, 2d - 1)$.*

Proof. Let $\hat{n} = \hat{k} + 1$, $\hat{m} = \hat{k} - \hat{d}$ so that $\chi_{\hat{d},\hat{k}}(z) = z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i$, and let $\hat{q} = \gcd(\hat{m}, \hat{n})$. Under the assumptions of the lemma, $\Psi_{d,\infty}(z) = \chi_{d,\infty}(z) = z^{d+1} - z^d - 1$, and $\Psi_{\hat{d},\hat{k}}(z) = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}$.

If $C(d, \infty) = C(\hat{d}, \hat{k})$, then by Theorem 15, $\Psi_{d,\infty}(z) = \Psi_{\hat{d},\hat{k}}(z)$, i.e.,

$$z^{d+1} - z^d - 1 = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}. \tag{22}$$

Now, note that $\xi_1(\Psi_{d,\infty}) = d + 1$, while $\xi_1(\Psi_{\hat{d},\hat{k}}) = \hat{n} - \hat{q} + 1$. Equating these, we get

$$d = \hat{n} - \hat{q}. \tag{23}$$

Next, observe that $\epsilon_1(\Psi_{d,\infty}) = d$. Additionally, we claim that $\epsilon_1(\Psi_{\hat{d},\hat{k}}) = \hat{q}$. This is because the smallest $k > 0$ such that the coefficient of z^k in $-\sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}$ is nonzero is precisely $\hat{q}$, and the term $-z^{\hat{q}}$ cannot be cancelled out by any term in $\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1}$. The reason that $-z^{\hat{q}}$ cannot get cancelled out is that the smallest exponent of z in $\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1}$ is $\hat{m}+1$, which is larger than $\hat{q}$, since $\hat{q} = \gcd(\hat{m}, \hat{n})$. Therefore, equating $\epsilon_1(\Psi_{d,\infty})$ and $\epsilon_1(\Psi_{\hat{d},\hat{k}})$, we get

$$d = \hat{q}. \tag{24}$$

From (23) and (24), we see that $\hat{n} = 2d$. Plugging this and $\hat{q} = d$ into (22), we get $z^{d+1} - z^d - 1 = \sum_{l=\frac{\hat{m}}{d}}^{1} z^{ld+1} - z^d - 1$. It follows that $\hat{m} = d$, and since $(\hat{m}, \hat{n}) = (\hat{k} - \hat{d}, \hat{k} + 1)$ by definition, the fact that $(\hat{m}, \hat{n}) = (d, 2d)$ implies that $(\hat{d}, \hat{k}) = (d-1, 2d-1)$. □

The proof of the above lemma involves arguments typical of those used in the proofs to follow. One especially important fact used in the above proof that should be kept in mind is that the function ϵ_1, when applied to the polynomial $\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}$, yields $\hat{q}$. Also, in all that is to follow, we shall continue to take $(\hat{m}, \hat{n})$ to be $(\hat{k} - \hat{d}, \hat{k} + 1)$ and $\hat{q}$ to be $\gcd(\hat{m}, \hat{n})$.

LEMMA 20. *If* $d \not\equiv 4 \pmod 6$ *and* $\hat{d}, \hat{k}$ *are such that* $\chi_{\hat{d},\hat{k}}(z)$ *is of Type* II, *then* $C(d, \infty) \neq C(\hat{d}, \hat{k})$.

Proof. With $d, \hat{d}, \hat{k}$ as in the statement of the lemma, we have $\Psi_{d,\infty}(z) = z^{d+1} - z^d - 1$, and

$$\Psi_{\hat{d},\hat{k}}(z) = \frac{\chi_{\hat{d},\hat{k}}(z)}{\Phi_{\hat{d},\hat{k}}(z)} = \frac{z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i}{(z^{\hat{r}}+1)\sum_{i=0}^{\hat{q}-1} z^i} = \frac{\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}}{z^{\hat{r}}+1},$$

where $\hat{r} = \gcd(\frac{\hat{m}}{2}+1, \hat{n}+1)$, and the last equality above comes from (21).

Suppose that $C(d, \infty) = C(\hat{d}, \hat{k})$, so that $\Psi_{d,\infty}(z) = \Psi_{\hat{d},\hat{k}}(z)$. Since the Ψ's are as given above, we have $(z^{\hat{r}}+1)(z^{d+1} - z^d - 1) = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}$, which upon expanding out the LHS becomes

$$z^{d+\hat{r}+1} + z^{d+1} - z^{d+\hat{r}} - z^{\hat{r}} - z^d - 1 = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}. \tag{25}$$

Our goal is to show that such an equality cannot arise for any $d, \hat{d}, \hat{k}$ satisfying the hypothesis of the lemma, leading to a contradiction that proves the lemma.

Applying ξ_1 to both sides of (25), we get $d + \hat{r} + 1 = \hat{n} - \hat{q} + 1$, implying

$$d + \hat{r} = \hat{n} - \hat{q}. \tag{26}$$

Next, note that the function ϵ_1, when applied to the RHS of (25), yields $\hat{q}$ and, when applied to the LHS, yields either d or $\hat{r}$, depending on whether $d \leq \hat{r}$ or $d > \hat{r}$. So, if $d \leq \hat{r}$, then $d = \hat{q}$, and if $d > \hat{r}$, then $\hat{q} = \hat{r}$. However, we cannot have $d > \hat{r}$, since $\hat{q} = \hat{r}$ is ruled out by Lemma 17.

Thus, we see that $d \le \hat{r}$, so that $d = \hat{q}$. Plugging this into (26), we get $\frac{\hat{n}}{\hat{q}} = 2 + \frac{\hat{r}}{d}$. Using this and $d = \hat{q}$, the RHS of (25) becomes

$$\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{r}}{d}+1} z^{ld+1} - \sum_{l=0}^{\frac{\hat{r}}{d}+1} z^{ld} = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{r}}{d}} z^{ld+1} - \sum_{l=0}^{\frac{\hat{r}}{d}} z^{ld} + z^{d+\hat{r}+1} - z^{d+\hat{r}}.$$

Therefore, upon cancelling out some terms common to both sides, (25) simplifies to $z^{d+1} - z^{\hat{r}} - z^d - 1 = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{r}}{d}} z^{ld+1} - \sum_{l=0}^{\frac{\hat{r}}{d}} z^{ld}$. Applying ξ_1 to both sides of this equality, we get $d + 1 = \hat{r} + 1$, i.e., $d = \hat{r}$. We thus have $\hat{q} = d = \hat{r}$, which is impossible by Lemma 17. □

LEMMA 21. *If* $d \not\equiv 4 \pmod 6$ *and* $\hat{d}, \hat{k}$ *are such that* $\chi_{\hat{d},\hat{k}}(z)$ *is of Type* III*, then* $C(d,\infty) \ne C(\hat{d},\hat{k})$.

Proof. An argument similar to that at the beginning of the proof of Lemma 20 shows that if $C(d,\infty) = C(\hat{d},\hat{k})$, with $d, \hat{d}, \hat{k}$ as above, then

$$\left(\sum_{i=0}^{\hat{r}-1} (-z)^i\right)(z^{d+1} - z^d - 1) = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}.$$

Equivalently, multiplying both sides by $z+1$, we have

$$(z^{\hat{r}} + 1)(z^{d+1} - z^d - 1) = (z+1)\left(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}\right).$$

Expanding out both sides of the above equation, we get

$$z^{d+\hat{r}+1} + z^{d+1} - z^{d+\hat{r}} - z^{\hat{r}} - z^d - 1 = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+2} - \sum_{l=0}^{\frac{\hat{m}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}. \tag{27}$$

Now, by definition of Type III, $\hat{r} \ge 3$, so that the term $-z^{d+\hat{r}}$ on the LHS of the above equation cannot get cancelled out by another term on the LHS. Therefore, the RHS must also have a $-z^{d+\hat{r}}$ term, and due to the negative sign, it must be one of the terms in $-\sum_{l=0}^{\frac{\hat{m}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}$. In other words, $d + \hat{r}$ must be one of the exponents of z in these two summations. Observe that the maximum exponent of z in these summations is $\max(\hat{m} - \hat{q} + 1, \hat{n} - \hat{q}) = \max(\hat{m}+1, \hat{n}) - \hat{q} = \hat{n} - \hat{q}$, since $\hat{n} > \hat{m}$. Therefore, $d + \hat{r} \le \hat{n} - \hat{q}$.

However, if we apply ξ_1 to both sides of (27), we find that $d + \hat{r} + 1 = \hat{n} - \hat{q} + 2$, so that $d + \hat{r} = \hat{n} - \hat{q} + 1$, which contradicts $d + \hat{r} \le \hat{n} - \hat{q}$. So, (27) cannot hold under the assumptions of the lemma, implying that $C(d,\infty)$ cannot be equal to $C(\hat{d},\hat{k})$. □

The last three lemmas show that when $d \not\equiv 4 \pmod 6$, then $C(d,\infty) = C(\hat{d},\hat{k})$ only if $(\hat{d},\hat{k}) = (d-1, 2d-1)$. The next three lemmas consider the case when $d \equiv 4 \pmod 6$. Recall that for any such d, $\Psi_{d,\infty}(z)$ is as given in (7).

LEMMA 22. *Let* $d \equiv 4 \pmod 6$ *and* $\hat{d}, \hat{k}$ *be such that* $\chi_{\hat{d},\hat{k}}(z)$ *is of Type* I. *Then,* $C(d,\infty) = C(\hat{d},\hat{k})$ *only if* $d = 4$ *and* $(\hat{d},\hat{k}) = (1,2)$.

Proof. If $C(d,\infty) = C(\hat{d},\hat{k})$ with $d, \hat{d}, \hat{k}$ as above, then we have

$$z^3 - z - 1 + \sum_{l=2}^{(d+2)/6} (z^{6l-3} - z^{6l-5} - z^{6l-6} + z^{6l-8}) = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}. \tag{28}$$

Applying ϵ_1 to both sides of this equation, we get $1 = \hat{q}$. Therefore, $\Phi_{\hat{d},\hat{k}}(z) = \sum_{i=0}^{\hat{q}-1} z^i = 1$, and hence $\Psi_{\hat{d},\hat{k}}(z) = \chi_{\hat{d},\hat{k}}(z)$. Thus, we must have $\Psi_{d,\infty}(z) = \chi_{\hat{d},\hat{k}}(z)$.

Now, the polynomial on the LHS of (28) can be of the form $z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i$ only if $d = 4$, since in this case it has no terms of the form $z^{6l-3} - z^{6l-5} - z^{6l-6} + z^{6l-8}$. So, $\Psi_{d,\infty}(z) = \chi_{\hat{d},\hat{k}}(z)$ implies that $d = 4$, in which case $\Psi_{d,\infty} = z^3 - z - 1 = \chi_{1,2}(z)$. Hence, $(\hat{d},\hat{k}) = (1,2)$, which proves the lemma. □

For the proofs of the next couple of lemmas, it is convenient to introduce the following notation: we shall use $\Omega(z^k)$ to denote an arbitrary polynomial of the form $\sum_{i=k}^{l} c_i z^i$, with $l \geq k$.

LEMMA 23. *Let $d \equiv 4 \pmod 6$ and $\hat{d}, \hat{k}$ be such that $\chi_{\hat{d},\hat{k}}(z)$ is of Type* II. *Then, $C(d,\infty) = C(\hat{d},\hat{k})$ only if $d = 4$ and $(\hat{d},\hat{k}) = (2,4)$.*

Proof. Arguing as in the proof of Lemma 20, we find that for the above choice of $d, \hat{d}, \hat{k}$, $C(d,\infty) = C(\hat{d},\hat{k})$ implies

$$(z^{\hat{r}} + 1)\left(z^3 - z - 1 + \sum_{l=2}^{(d+2)/6} (z^{6l-3} - z^{6l-5} - z^{6l-6} + z^{6l-8})\right) = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}.$$

As usual, we now apply ϵ_1 to both sides of this equation, which yields $1 = \hat{q}$. Hence $\Phi_{\hat{d},\hat{k}}(z) = (z^{\hat{r}} + 1)\sum_{i=0}^{\hat{q}-1} z^i = z^{\hat{r}} + 1$. Therefore, $\chi_{\hat{d},\hat{k}}(z) = \Phi_{\hat{d},\hat{k}}(z)\Psi_{\hat{d},\hat{k}}(z) = \Phi_{\hat{d},\hat{k}}(z)\Psi_{d,\infty}(z)$, which shows that

$$\chi_{\hat{d},\hat{k}}(z) = (z^{\hat{r}} + 1)\left(z^3 - z - 1 + \sum_{l=2}^{(d+2)/6} (z^{6l-3} - z^{6l-5} - z^{6l-6} + z^{6l-8})\right). \tag{29}$$

Note that since $\hat{q} = 1$, by Lemma 17, $\hat{r} \geq 2$.

Suppose first that $d = 4$ so that $\Psi_{d,\infty}(z) = z^3 - z - 1$. Then, (29) becomes $\chi_{\hat{d},\hat{k}}(z) = (z^{\hat{r}} + 1)(z^3 - z - 1)$, which is the same as

$$\chi_{\hat{d},\hat{k}}(z) = z^{\hat{r}+3} + z^3 - z^{\hat{r}+1} - z^{\hat{r}} - z - 1. \tag{30}$$

Since only the leading coefficient of the polynomial $\chi_{\hat{d},\hat{k}}(z)$ is positive, either $z^{\hat{r}+3}$ or z^3 must be eliminated by one of the other terms on the RHS of (30). As $\hat{r} + 3$ is strictly larger than any other exponent of z on the RHS, z^3 is the term that must get eliminated, and this can happen only if either $\hat{r} = 3$ or $\hat{r} + 1 = 3$, i.e., $\hat{r} = 2$. If $\hat{r} = 3$, then the RHS of (30) turns out to be $z^6 - z^4 - z - 1$, which is not of the form $z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i$. So, we must have $\hat{r} = 2$, in which case the RHS of (30) becomes $z^5 - z^2 - z - 1 = \chi_{2,4}(z)$. So, one possible solution for $C(d,\infty) = C(\hat{d},\hat{k})$ is $(d,\hat{d},\hat{k}) = (4,2,4)$.

Now, suppose that $d > 4$, so that $d \geq 10$, as 10 is the next largest integer that is equivalent to 4 (mod 6). Then, $\Psi_{d,\infty}(z) = -1 - z + z^3 + z^4 + \Omega(z^5)$, and (29) becomes

$$\chi_{\hat{d},\hat{k}}(z) = z^{\hat{r}+4} + z^{\hat{r}+3} + z^4 + z^3 - z^{\hat{r}+1} - z^{\hat{r}} - z - 1 + \Omega(z^5). \tag{31}$$

Note that if $\hat{r} \geq 5$, then the RHS above becomes $z^4 + z^3 - z - 1 + \Omega(z^5)$, which cannot be of the form $z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i$. So, we must have $\hat{r} = 2$, 3 or 4.

If $\hat{r} = 2$, then the RHS of (29) is of the form $z^{d+\hat{r}-1} + \sum_{i=5}^{d+\hat{r}-2} c_i z^i + z^4 - z^2 - z - 1$, which cannot be $\chi_{\hat{d},\hat{k}}(z)$ for any $\hat{d}, \hat{k}$. Similarly, if $\hat{r} = 4$, then the RHS of (29) is of the form $z^{d+\hat{r}-1} + \sum_{i=5}^{d+\hat{r}-2} c_i z^i + z^3 - z - 1$, which cannot be any $\chi_{\hat{d},\hat{k}}(z)$.

Finally, if $\hat{r} = 3$, then the RHS of (29) becomes $z^{d+\hat{r}-1} + \sum_{i=5}^{d+\hat{r}-2} c_i z^i - z - 1$, which can at best be $z^{d+\hat{r}-1} - z - 1 = z^{d+2} - z - 1 = \chi_{d,d+1}(z)$. But this too does not yield a solution to $C(d,\infty) = C(\hat{d},\hat{k})$, since it is clear that $C(d,\infty) \neq C(d, d+1)$ for any d. This completes the analysis of the $d > 4$ case and hence the proof of the lemma. ∎

Lemma 24. *Let* $d \equiv 4 \pmod 6$ *and* $\hat{d}, \hat{k}$ *be such that* $\chi_{\hat{d},\hat{k}}(z)$ *is of Type* III. *Then,* $C(d,\infty) = C(\hat{d},\hat{k})$ *only if* $(\hat{d},\hat{k}) = (d-1, 2d-1)$.

Proof. With $d, \hat{d}, \hat{k}$ as in the above statement, if $C(d,\infty) = C(\hat{d},\hat{k})$, then the usual argument shows that we must have

$$\left(\sum_{i=0}^{\hat{r}-1} (-z)^i\right) \Psi_{d,\infty}(z) = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}} \tag{32}$$

with $\Psi_{d,\infty}(z)$ having the form given in (7).

Note that as $\chi_{\hat{d},\hat{k}}(z)$ is of Type III, we must have $\hat{r} \geq 3$ odd. Suppose first that $\hat{r} = 3$. Then the LHS of the above equation is $(z^2 - z + 1)\Psi_{d,\infty}(z) = \chi_{d,\infty}(z) = z^{d+1} - z^d - 1$ by Theorem 3. Therefore, (32) in this case is identical to (22) in the proof of Lemma 19. As analyzed there, this equation implies $(\hat{d},\hat{k}) = (d-1, 2d-1)$.

So, we are left with the case $\hat{r} \geq 5$. Note that the LHS of (32) may be written as

$$\begin{aligned}\left(z^2 - z + 1 + \sum_{i=3}^{\hat{r}-1} (-z)^i\right) \Psi_{d,\infty}(z) &= (z^2 - z + 1)\Psi_{d,\infty}(z) - z^3 \left(\sum_{i=0}^{\hat{r}-4} (-z)^i\right) \Psi_{d,\infty}(z) \\ &= z^{d+1} - z^d - 1 - z^3 \left(\sum_{i=0}^{\hat{r}-4} (-z)^i\right) \Psi_{d,\infty}(z).\end{aligned}$$

Therefore, if we multiply both sides of (32) by $\sum_{i=0}^{\hat{q}-1} z^i$ and use (21), then the resulting equation can be written as

$$\begin{aligned}\chi_{\hat{d},\hat{k}}(z) &= (z^{d+1} - z^d - 1) \sum_{i=0}^{\hat{q}-1} z^i - z^3 \left(\sum_{i=0}^{\hat{r}-4} (-z)^i\right) \Psi_{d,\infty}(z) \sum_{i=0}^{\hat{q}-1} z^i \\ &= z^{d+\hat{q}} - z^d - \sum_{i=0}^{\hat{q}-1} z^i + z^3 + \Omega(z^4),\end{aligned} \tag{33}$$

where we have used the fact that $(z^{d+1} - z^d - 1) \sum_{i=0}^{\hat{q}-1} z^i = z^{d+\hat{q}} - z^d - \sum_{i=0}^{\hat{q}-1} z^i$.

Now, the fact that $\hat{r} = \gcd(\frac{\hat{m}}{2} + 1, \hat{n} + 1) \geq 5$ implies that $\frac{\hat{m}}{2} + 1 \geq 5$, which means that $\hat{m} \geq 8$. Therefore, $\chi_{\hat{d},\hat{k}}(z) = z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i$ must contain the sequence $-z^8 - z^7 - \cdots - z - 1$. In particular, the coefficient of z^3 in $\chi_{\hat{d},\hat{k}}(z)$ is -1. However, on the RHS of (33), there are at most two z^3 terms, one of which is $+z^3$, and the other is $-z^3$ from the summation $-\sum_{i=0}^{\hat{q}-1} z^i$ if $\hat{q} - 1 \geq 3$. So, the coefficient of z^3

on the RHS of (33) can either be 0 or +1, which implies that the RHS cannot be of the form required by $\chi_{\hat{d},\hat{k}}(z)$. Therefore, we cannot have $C(d,\infty) = C(\hat{d},\hat{k})$ when $\hat{r} \geq 5$. □

Lemmas 19–24 together prove the following result, which is the part of Theorem 1 dealing with the case when one of the (d,k) constraints involved is a (d,∞) constraint.

THEOREM 25. *If $d, \hat{d}, \hat{k}$ are nonnegative integers such that $C(d,\infty) = C(\hat{d},\hat{k})$, then one of the following holds:*

(i) $(\hat{d},\hat{k}) = (d-1, 2d-1)$,

(ii) $d = 4$ *and* $(\hat{d},\hat{k})$ *is either* $(1,2)$ *or* $(2,4)$.

We now move on to analyze the equality $C(d,k) = C(\hat{d},\hat{k})$ when $k, \hat{k}$ are both finite. Once again, we perform a case-by-case analysis of the various situations that arise when each of the characteristic polynomials involved is of one of the three types defined earlier. Because of symmetry, there are only six cases to be considered—three when $\chi_{d,k}(z)$ and $\chi_{\hat{d},\hat{k}}(z)$ are of the same type and three more as follows: (a) $\chi_{d,k}(z)$ of Type I, $\chi_{\hat{d},\hat{k}}(z)$ of Type II, (b) $\chi_{d,k}(z)$ of Type I, $\chi_{\hat{d},\hat{k}}(z)$ of Type III, and (c) $\chi_{d,k}(z)$ of Type II, $\chi_{\hat{d},\hat{k}}(z)$ of Type III.

The situation when $\chi_{d,k}(z)$ and $\chi_{\hat{d},\hat{k}}(z)$ are both of Type I is the easiest to deal with, and we dispose of this first. As usual, we define $(m,n) = (k-d, k+1)$, $(\hat{m},\hat{n}) = (\hat{k}-\hat{d}, \hat{k}+1)$, $q = \gcd(m,n)$, and $\hat{q} = \gcd(\hat{m},\hat{n})$.

LEMMA 26. *Let $d, k, \hat{d}, \hat{k}$ be such that $\chi_{d,k}(z)$ and $\chi_{\hat{d},\hat{k}}(z)$ are both of Type* I. *Then, $C(d,k) = C(\hat{d},\hat{k})$ only if $(d,k) = (\hat{d},\hat{k})$.*

Proof. By Theorem 15, $C(d,k) = C(\hat{d},\hat{k})$ implies $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$. Since $\chi_{d,k}(z)$ and $\chi_{\hat{d},\hat{k}}(z)$ are both of Type I, we have an explicit form for their nonreciprocal parts, using which we get

$$\sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq} = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}. \tag{34}$$

Applying ϵ_1 to both sides of the above equation, we get $q = \hat{q}$. But this means that $\Phi_{d,k}(z) = \sum_{i=0}^{q-1} z^i = \sum_{i=0}^{\hat{q}-1} z^i = \Phi_{\hat{d},\hat{k}}(z)$. Thus, the polynomials $\chi_{d,k}(z)$ and $\chi_{\hat{d},\hat{k}}(z)$ have identical reciprocal parts and identical nonreciprocal parts, which shows that $\chi_{d,k}(z) = \chi_{\hat{d},\hat{k}}(z)$, i.e., $(d,k) = (\hat{d},\hat{k})$. □

When $\chi_{d,k}(z)$ and $\chi_{\hat{d},\hat{k}}(z)$ are both of Type II or Type III, the analysis involves the use of the following technical lemma, whose proof we defer to the end of this paper.

LEMMA 27. *Let $m, n, r, \hat{m}, \hat{n}, \hat{r}$ be positive integers such that $n > m$ and $\hat{n} > \hat{m}$. If $(z^r+1)\left(z^n - \sum_{i=0}^{m} z^i\right) = (z^{\hat{r}}+1)(z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i)$, then $(m,n,r) = (\hat{m},\hat{n},\hat{r})$.*

In all that is to follow, we shall take $r = \gcd(\frac{m}{2}+1, n+1)$ and $\hat{r} = \gcd(\frac{\hat{m}}{2}+1, \hat{n}+1)$, whenever $m, \hat{m}$ are even.

LEMMA 28. *Let $d, k, \hat{d}, \hat{k}$ be such that $\chi_{d,k}(z)$ and $\chi_{\hat{d},\hat{k}}(z)$ are either both of Type* II *or both of Type* III. *Then, $C(d,k) = C(\hat{d},\hat{k})$ only if $(d,k) = (\hat{d},\hat{k})$.*

Proof. Suppose first that $\chi_{d,k}(z)$ and $\chi_{\hat{d},\hat{k}}(z)$ are both of Type II. As shown in the proof of Lemma 20, we have

$$\Psi_{d,k}(z) = \frac{\sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq}}{z^r+1}, \qquad \Psi_{\hat{d},\hat{k}}(z) = \frac{\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}}{z^{\hat{r}}+1}.$$

Therefore, if $C(d,k) = C(\hat{d},\hat{k})$, then we have $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$, from which it follows that

$$(z^{\hat{r}}+1)\left(\sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq}\right) = (z^{r}+1)\left(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}\right). \tag{35}$$

We shall consider the following four cases individually: (i) $r \le \hat{q}$ and $q < \hat{r}$, (ii) $r \le \hat{q}$ and $q \ge \hat{r}$, (iii) $r > \hat{q}$ and $q < \hat{r}$, and (iv) $r > \hat{q}$ and $q \ge \hat{r}$. Observe, however, that (iv) is the same as (i), with the roles of (d,k) and $(\hat{d},\hat{k})$ reversed. So, it suffices to consider the first three cases only.

We consider case (i) first. In this case, applying ϵ_1 to both sides of (35), we find that $q = r$, which is impossible by Lemma 17.

In case (ii), applying ϵ_1 to both sides of (35) yields $r = \hat{r}$. Hence (35) reduces to (34) in the proof of Lemma 26, which as shown in that proof, leads to the conclusion that $(d,k) = (\hat{d},\hat{k})$.

Moving on to case (iii), applying ϵ_1 to (35) here yields $q = \hat{q}$. Hence multiplying both sides of (35) by $\sum_{i=0}^{q-1} z^i$, we get via (21) $(z^{\hat{r}}+1)\left(z^n - \sum_{i=0}^{m} z^i\right) = (z^r+1)(z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i)$. But now Lemma 27 shows that $(m,n) = (\hat{m},\hat{n})$, which implies that $(d,k) = (\hat{d},\hat{k})$ in this case as well. Thus, we have shown that when $\chi_{d,k}(z)$ and $\chi_{\hat{d},\hat{k}}(z)$ are both of Type II, then $C(d,k) = C(\hat{d},\hat{k})$ is possible only if $(d,k) = (\hat{d},\hat{k})$.

If $\chi_{d,k}(z), \chi_{\hat{d},\hat{k}}(z)$ are both of Type III, then using (21), we find that

$$\Psi_{d,k}(z) = \frac{\sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq}}{\sum_{i=0}^{r-1}(-z)^i}, \qquad \Psi_{\hat{d},\hat{k}}(z) = \frac{\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}}{\sum_{i=0}^{\hat{r}-1}(-z)^i}.$$

So, from $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$, we obtain

$$\left(\sum_{i=0}^{\hat{r}-1}(-z)^i\right)\left(\sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq}\right) = \left(\sum_{i=0}^{r-1}(-z)^i\right)\left(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}\right).$$

Multiplying both sides of the above equation by $z+1$, we obtain (35), which as shown above, leads to $(d,k) = (\hat{d},\hat{k})$. □

At this point, we would like to remark that Lemmas 26 and 28 actually prove the following interesting fact: if two polynomials of the same type (I, II, or III) have identical nonreciprocal parts, then the polynomials themselves are identical. In other words, within each of the three type classes, a polynomial is uniquely determined by its nonreciprocal part.

We are now only left to deal with the three cases where the characteristic polynomials are of different types. The next three lemmas consider each case in turn.

LEMMA 29. *Let $d,k,\hat{d},\hat{k}$ be such that $\chi_{d,k}(z)$ is of Type* I *and $\chi_{\hat{d},\hat{k}}(z)$ is of Type* II*. Then, $C(d,k) = C(\hat{d},\hat{k})$ only if $(d,k) = (d,2d)$ and $(\hat{d},\hat{k}) = (d+1,3d+1)$.*

Proof. With $d,k,\hat{d},\hat{k}$ as above, we have

$$\Psi_{d,k}(z) = \sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq}, \qquad \Psi_{\hat{d},\hat{k}}(z) = \frac{\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}}{z^{\hat{r}}+1}.$$

So, if $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$, then it follows that

$$(z^{\hat{r}}+1)\left(\sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq}\right) = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}. \tag{36}$$

Note that if $\hat{r} \leq q$, then applying ϵ_1 to both sides of (36), we get $\hat{r} = \hat{q}$, which is impossible by Lemma 17. Hence, we must have $\hat{r} > q$.

So, applying ϵ_1 to (36) yields $q = \hat{q}$. Therefore, multiplying both sides of (36) by $\sum_{i=0}^{q-1} z^i$, we obtain, on account of (21), $(z^{\hat{r}}+1)\left(z^n - \sum_{i=0}^{m} z^i\right) = z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i$ or, equivalently,

$$z^{n+\hat{r}} + z^n - \sum_{i=0}^{m} z^{\hat{r}+i} - \sum_{i=0}^{m} z^i = z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i. \tag{37}$$

We claim that the equality in (37) is possible only if $\hat{r} = m+1$ and $n = 2m+1$, in which case the LHS of the equation is $\chi_{m+1,3m+1}(z)$. To prove this claim, we observe first that if $\hat{r} \leq m$, then on the LHS of (37), the coefficient of $z^{\hat{r}}$ is -2. This is because we have one $-z^{\hat{r}}$ term coming from the summation $-\sum_{i=0}^{m} z^{\hat{r}+i}$ and another from the summation $-\sum_{i=0}^{m} z^i$, and neither of these terms can be cancelled out by z^n or $z^{n+\hat{r}}$, since $n > m \geq \hat{r}$. However, since there cannot be any term with coefficient -2 on the RHS, we must have $\hat{r} > m$.

Also, $\hat{r} > m+1$ is impossible, since if this were the case, $z^n - \sum_{i=0}^{m} z^{\hat{r}+i} - \sum_{i=0}^{m} z^i$ cannot be of the form $-\sum_{i=0}^{\hat{m}} z^i$, as can be easily verified. Thus, we are forced to conclude that for (37) to hold, $\hat{r}$ must be equal to $m+1$.

With $\hat{r} = m+1$, the LHS of (37) becomes $z^{n+m+1} + z^n - \sum_{i=0}^{2m+1} z^i$, which can be of the form $z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i$ only if $n = 2m+1$, so that z^n cancels out with $-z^{2m+1}$. With this choice of $\hat{r}$ and m, the LHS of (37) reduces to $z^{3m+2} - \sum_{i=0}^{2m} z^i = \chi_{m+1,3m+1}(z)$. Hence we see that $(\hat{d},\hat{k}) = (m+1, 3m+1)$, and as $(m,n) = (m, 2m+1)$, we also have $(d,k) = (m, 2m)$, which proves the lemma. □

Lemma 30. *Let $d,k,\hat{d},\hat{k}$ be such that $\chi_{d,k}(z)$ is of Type* I *and $\chi_{\hat{d},\hat{k}}(z)$ is of Type* III. *Then, $C(d,k) = C(\hat{d},\hat{k})$ only if $(d,k) = (d,2d)$ and $(\hat{d},\hat{k}) = (d+1, 3d+1)$, or $(d,k) = (1,2)$ and $(\hat{d},\hat{k}) = (3,7)$.*

Proof. For the above choice of $d,k,\hat{d},\hat{k}$, it follows from $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$ that $(\sum_{i=0}^{\hat{r}-1}(-z)^i)(\sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq}) = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}$, which upon multiplying by $z+1$ becomes

$$(z^{\hat{r}}+1)\left(\sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq}\right) = (z+1)\left(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}\right). \tag{38}$$

Now, the RHS above can be written as $(z+1)(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=1}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}) - z - 1$. Note that the $-z$ term cannot get cancelled out by any other term, since the smallest exponent of z in $(z+1)\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1}$ is $\hat{m}+1 \geq 2$. Therefore, ϵ_1 applied to the RHS of (38) yields 1.

When ϵ_1 is applied to the LHS of (38), we either get $\hat{r}$ if $\hat{r} \leq q$, or we get q if $\hat{r} > q$. Therefore, either $\hat{r} = 1$ or $q = 1$. However, $\hat{r} = 1$ is impossible because $\hat{r} \geq 3$ by definition of Type III polynomials. Hence, we must have $q = 1$.

Therefore, (38) reduces to

$$(z^{\hat{r}} + 1)\left(z^n - \sum_{i=0}^{m} z^i\right) = (z+1)\left(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}\right). \tag{39}$$

We will show that if $m \geq 2$, then the above equality is possible only if $(d,k) = (d, 2d)$ and $(\hat{d}, \hat{k}) = (d+1, 3d+1)$, and if $m = 1$, then the equality above implies that $(d,k) = (1,2)$ and $(\hat{d}, \hat{k}) = (3,7)$.

So, suppose first that $m \geq 2$. The LHS of (39) can be written as $z^{n+\hat{r}} + z^n - \sum_{i=0}^{m} z^{\hat{r}+i} - \sum_{i=0}^{m} z^i$. Since $\hat{r} \geq 3$ and $m \geq 2$, the coefficient of z^2 in this polynomial is -1. Now, since $\hat{q} \geq 2$ by definition of Type III polynomials, there can be a $-z^2$ term on the RHS of (39) only if $\hat{q} = 2$. Therefore, it follows from (21) that the RHS of (39) is

$$(z+1)\,\frac{z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i}{\sum_{i=0}^{\hat{q}-1} z^i} = (z+1)\,\frac{z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i}{z+1} = z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i.$$

Thus, we see that when $m \geq 2$, we must have $\hat{q} = 2$, and furthermore, (39) reduces to (37). But, as shown in the proof of Lemma 29, (37) holds only if $(d,k) = (d, 2d)$ and $(\hat{d}, \hat{k}) = (d+1, 3d+1)$.

It only remains to consider the case when $m = 1$. In this case, the LHS of (39) is $(z^{\hat{r}} + 1)(z^n - z - 1) = z^{n+\hat{r}} + z^n - z^{\hat{r}+1} - z^{\hat{r}} - z - 1$. Cancelling out $-z-1$ from both sides of (39), we get

$$z^{n+\hat{r}} + z^n - z^{\hat{r}+1} - z^{\hat{r}} = (z+1)\left(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=1}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}\right). \tag{40}$$

Now, ϵ_1 applied to the RHS above yields $\hat{q}$, and the coefficient of $z^{\hat{q}}$ is -1. The $-z^{\hat{q}}$ term on the RHS must correspond to either $-z^{\hat{r}}$ or $-z^{\hat{r}+1}$ on the LHS. Since $\hat{q} \neq \hat{r}$ by Lemma 17, the $-z^{\hat{q}}$ term on the RHS must correspond to the $-z^{\hat{r}+1}$ term on the LHS, showing that $\hat{q} = \hat{r} + 1$. Therefore, ϵ_1 when applied to the LHS of (40) must yield $\hat{r} + 1$, which means that $-z^{\hat{r}}$ must get cancelled by z^n so that we must have $n = \hat{r}$. Finally, applying ξ_1 to (40), we also obtain $n + \hat{r} = \hat{n} - \hat{q} + 2$. Using $\hat{q} = \hat{r} + 1$ and $n = \hat{r}$ to eliminate $\hat{q}$ and $\hat{r}$ from this last equation, we get $\hat{n} = 3n - 1$.

Since $\hat{q} = \hat{r} + 1 = n + 1$ and $\hat{q} | \hat{n}$, we find that $n + 1$ must divide $3n - 1$. Writing $3n - 1$ as $3(n+1) - 4$, we see that $n+1$ must be a factor of 4. Hence $n = 0, 1$, or 3. But as $n > m \geq 1$, n must in fact be 3. Hence $\hat{n} = 3n - 1 = 8$. Furthermore, $\hat{q} = n + 1 = 4$, and so the facts that $\hat{q} | \hat{m}$ and $\hat{m} < \hat{n}$ now imply that $\hat{m} = 4$. Thus, we have shown that when $m = 1$, equality in (38) is possible only if $n = 3$ and $(\hat{m}, \hat{n}) = (4, 8)$. As these values of (m,n) and $(\hat{m}, \hat{n})$ are equivalent to $(d,k) = (1,2)$ and $(\hat{d}, \hat{k}) = (3,7)$, the proof of the lemma is complete. □

LEMMA 31. *Let $d, k, \hat{d}, \hat{k}$ be such that $\chi_{d,k}(z)$ is of Type* II *and $\chi_{\hat{d},\hat{k}}(z)$ is of Type* III. *Then, $C(d,k) = C(\hat{d}, \hat{k})$ only if $(d,k) = (d, 2d)$ and $(\hat{d}, \hat{k}) = (d+1, 3d+1)$.*

Proof. When $\chi_{d,k}(z)$ is of Type II and $\chi_{\hat{d},\hat{k}}(z)$ is of Type III, from the equality $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$ we get, via (21),

$$\left(\sum_{i=0}^{\hat{r}-1}(-z)^i\right)\left(\sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq}\right) = (z^r+1)\left(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}\right).$$

Upon multiplying both sides of this equation by $z+1$, we obtain

$$(z^{\hat{r}}+1)\left(\sum_{l=\frac{m}{q}}^{\frac{n}{q}-1} z^{lq+1} - \sum_{l=0}^{\frac{n}{q}-1} z^{lq}\right) = (z+1)(z^r+1)\left(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}\right). \tag{41}$$

Using $(z^r+1)(z+1) = z^{r+1}+z^r+z+1$, we can write the RHS above as

$$(z+1)(z^r+1)\left(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=1}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}\right) - (z^{r+1}+z^r+z+1). \tag{42}$$

Recall that the definition of Type III requires $\hat{q} \geq 2$ even and $\hat{r} \geq 3$ odd. Since $\hat{m} \geq \hat{q} \geq 2$, the smallest exponent of z in $(z+1)(z^r+1)(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1})$ is $\hat{m}+1 \geq 3$. Hence the $-z$ term in (42) cannot be cancelled out by any other term. It follows that the coefficient of z on the RHS of (41) is nonzero, and so this must be true on the LHS as well. But, the only way for the coefficient of z to be nonzero on the LHS is if $\hat{r} = 1$ or $q = 1$. The former is impossible since $\hat{r} \geq 3$. So, we must have $q = 1$, and consequently the LHS of (41) simplifies to $(z^{\hat{r}}+1)\left(z^n - \sum_{i=0}^{m} z^i\right)$. Expanding out the product $(z+1)(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}})$ on the RHS of (41), we can rewrite (41) as

$$(z^{\hat{r}}+1)\left(z^n - \sum_{i=0}^{m} z^i\right) = (z^r+1)\left(\sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+2} - \sum_{l=0}^{\frac{\hat{m}}{\hat{q}}-1} z^{l\hat{q}+1} - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}\right). \tag{43}$$

We have thus far shown that for $\Psi_{d,k}(z) = \Psi_{\hat{d},\hat{k}}(z)$ to be true for $d, k, \hat{d}, \hat{k}$ as in the statement of the lemma, then we must have $q = 1$ and (43) must hold. Our aim now is to show that for $q = 1$ and (43) to be true, we must also have $r = 2$ and $\hat{q} = 4$, from which it will follow that $(d,k) = (d, 2d)$ and $(\hat{d}, \hat{k}) = (d+1, 3d+1)$.

The first step in this process is to show that $\hat{q} \neq 2$ so that (since $\hat{q}$ is even) $\hat{q} \geq 4$. If we assume that $\hat{q} = 2$, then it is easily seen that the RHS of (43) simplifies to $(z^r+1)(z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i)$. Therefore, by Lemma 27, (43) holds only if $(m,n) = (\hat{m}, \hat{n})$, or equivalently $(d,k) = (\hat{d}, \hat{k})$, which cannot happen since $\chi_{d,k}(z)$ and $\chi_{\hat{d},\hat{k}}(z)$ are of different types. Hence, $\hat{q} = 2$ is impossible, and so $\hat{q} \geq 4$. We next show that this, along with the fact that $q = 1$, implies that $r = 2$.

Note that m is even, for if it were odd, then by Theorem 11, $\chi_{d,k}(z)$ would be of Type I. Hence $m \geq 2$, from which it follows that the LHS of (43) contains a $-z^2$ term, i.e., the coefficient of z^2 on the LHS is -1. Therefore, the RHS of (43) must also contain a $-z^2$ term, which since $\hat{q} \geq 4$, can happen only if $r = 1$ or 2. But since $q = 1$, Lemma 17 forces r to be 2.

Setting $r = 2$, it can be verified that (43), upon multiplying out the product on its RHS, becomes

$$(z^{\hat{r}} + 1)\left(z^n - \sum_{i=0}^{m} z^i\right) = \sum_{l=\frac{\hat{m}}{\hat{q}}}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}+4} - \sum_{l=0}^{\frac{\hat{m}}{\hat{q}}-1} (z^{l\hat{q}+3} + z^{l\hat{q}+2} + z^{l\hat{q}+1}) - \sum_{l=0}^{\frac{\hat{n}}{\hat{q}}-1} z^{l\hat{q}}. \tag{44}$$

We now show that the above equality can hold only if $\hat{q} = 4$. Suppose, to the contrary, that $\hat{q} \neq 4$, so that $\hat{q} \geq 6$. Observe that since $\hat{q} \geq 6$, no cancellation of terms is possible among the various summations on the RHS of (44), as the exponents in different summations leave different remainders modulo $\hat{q}$. It follows that the RHS of (44) is of the form $-1 - z - z^2 - z^3 + \Omega(z^6)$, where $\Omega(z^6)$ denotes some polynomial of the form $\sum_{k \geq 6} c_k z^k$. In particular, the RHS cannot contain any z^4 or z^5 terms.

On the other hand, the LHS of (44) is $z^{n+\hat{r}} + z^n - \sum_{i=0}^{m} z^{\hat{r}+i} - \sum_{i=0}^{m} z^i$. Note that neither $z^{n+\hat{r}}$ nor z^n can cancel out any term in the summation $-\sum_{i=0}^{m} z^i$, so that all the terms in this summation remain intact on the LHS. But as the LHS cannot contain any z^4 or z^5 terms (because the RHS does not contain such terms), we find that $m \leq 3$. However, as observed earlier, m is even, so that we must in fact have $m = 2$. But now, in order for the LHS to contain a $-z^3$ term, we must either have $\hat{r} = 3$, or $n = \hat{r}$ and $\hat{r} + 1 = 3$. The latter is impossible as it implies that $n = 2 = m$, which cannot happen. But $\hat{r} = 3$ is also impossible, since with $\hat{r} = 3$ and $m = 2$, the LHS reduces to $z^{n+\hat{r}} + z^n - \sum_{i=0}^{5} z^i$, which will always contain a z^4 or z^5 term. Thus, if we assume that $\hat{q} \neq 4$, we are forced to conclude that (44) cannot hold.

Therefore, for (44) to hold, we must have $\hat{q} = 4$. But with $\hat{q} = 4$, it is readily verified that the RHS of (44) simplifies to $z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i$. As a result, (44) becomes identical to (37) in the proof of Lemma 29, and as shown there, equality in (37) is possible only if $(d, k) = (d, 2d)$ and $(\hat{d}, \hat{k}) = (d + 1, 3d + 1)$. This completes the proof of the lemma. □

Lemmas 26 and 28–31 together prove the following theorem, which in conjunction with Theorem 25 forms Theorem 1.

THEOREM 32. *If $d, k, \hat{d}, \hat{k}$ are nonnegative integers such that $C(d, k) = C(\hat{d}, \hat{k})$, but $(d, k) \neq (\hat{d}, \hat{k})$, then one of the following holds:*

(i) $\{(d, k), (\hat{d}, \hat{k})\} = \{(\ell, 2\ell), (\ell + 1, 3\ell + 1)\}$ *for some integer* $\ell \geq 0$.

(ii) $\{(d, k), (\hat{d}, \hat{k})\} = \{(1, 2), (3, 7)\}$.

There still remains a loose end that needs to be tied up, namely, a proof of Lemma 27. We provide such a proof now.

Proof of Lemma 27. Suppose that $m, n, r, \hat{m}, \hat{n}, \hat{r}$ are as in the statement of the lemma and that

$$(z^r + 1)\left(z^n - \sum_{i=0}^{m} z^i\right) = (z^{\hat{r}} + 1)\left(z^{\hat{n}} - \sum_{i=0}^{\hat{m}} z^i\right). \tag{45}$$

It suffices to show that $r = \hat{r}$.

Multiplying both sides of (45) by $z - 1$, we obtain

$$(z^r + 1)(z^{n+1} - z^n - z^{m+1} + 1) = (z^{\hat{r}} + 1)(z^{\hat{n}+1} - z^{\hat{n}} - z^{\hat{m}+1} + 1). \tag{46}$$

Observe first that upon comparing the degrees of both sides of the above equation, we get

$$n + r = \hat{n} + \hat{r}. \tag{47}$$

Taking the derivative of both sides of (46) and setting $z = 1$ yields $-2m = -2\hat{m}$, so that $m = \hat{m}$. Next, taking the second derivative of both sides of (46) and setting $z = 1$, we get

$$4n - 2m^2 - 2mr - 2m = 4\hat{n} - 2\hat{m}^2 - 2\hat{m}\hat{r} - 2\hat{m}.$$

Using the fact that $m = \hat{m}$, the above equation reduces to

$$4n - 2mr = 4\hat{n} - 2m\hat{r}. \tag{48}$$

But now, using (47) and (48), we have

$$(2m + 4)r = 4(r + n) - (4n - 2mr) = 4(\hat{r} + \hat{n}) - (4\hat{n} - 2m\hat{r}) = (2m + 4)\hat{r}.$$

Since $m \neq -2$, as $m > 0$, we must have $r = \hat{r}$, as desired. □

Acknowledgment. The authors would like to thank the anonymous reviewer for a thorough review of the paper, for simplifying a few of the proofs in the paper, and especially for the clever proof of Lemma 27.

REFERENCES

[1] M. Filaseta, *On the factorization of polynomials with small Euclidean norm*, in Number Theory in Progress, Vol. 1, de Gruyter, Berlin, 1999, pp. 143–163.

[2] K.A.S. Immink, P.H. Siegel and J.K. Wolf, *Codes for digital recorders*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2260–2299.

[3] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, Cambridge, UK, 1995.

[4] W. Ljunggren, *On the irreducibility of certain trinomials and quadrinomials*, Math. Scand., 8 (1960), pp. 65–70.

[5] B.H. Marcus, R.M. Roth, and P.H. Siegel, *Constrained systems and coding for recording channels*, in Handbook of Coding Theory, R. Brualdi, C. Huffman, and V. Pless, eds., Elsevier, Amsterdam, The Netherlands, 1998.

[6] W.H. Mills, *The factorization of certain quadrinomials*, Math. Scand., 57 (1985), pp. 44–50.

[7] E.S. Selmer, *On the irreducibility of certain trinomials*, Math. Scand., 4 (1956), pp. 287–302.

[8] C.E. Shannon, *A mathematical theory of communication*, Bell Syst. Tech. J., 27 (1948), pp. 379–423.

[9] D.A. Wolfram, *Solving generalized Fibonacci recurrences*, Fibonacci Quart., 36 (1998), pp. 129–145.

SIAM J. Discrete Math.
Vol. 17, No. 2, pp. 298–319

MAXIMIZING THE SHANNON CAPACITY OF CONSTRAINED SYSTEMS WITH TWO CONSTRAINTS[*]

NAVIN KASHYAP[†]

Abstract. In this paper, we consider the problem of finding the set $\{A, B\} \subset \{0,1\}^m$ that maximizes, among all 2-subsets of $\{0,1\}^m$, the Shannon capacity, $H(A,B)$, of a constrained system of binary sequences that do not contain A or B as a contiguous subsequence. This problem is motivated by the problem of finding a pair of length-m binary sequences, called markers, that achieves the maximum rate, $R(2,m,n)$, of a $(2,m,n)$ periodic prefix-synchronized (PPS) code. A $(2,m,n)$ PPS code is a binary block code with two length-m markers, A, B, and codewords of length n that inserts A and B alternately at regular intervals in the encoded bitstream, with the additional constraint that A and B may not appear anywhere in the encoded bitstream other than where inserted. We show that for any $m \geq 2$, $\lim_{n\to\infty} R(2,m,n) = \max\{H(A,B) : \{A,B\} \subset \{0,1\}^m\} = \log_2 \rho_{m-1}$, where ρ_{m-1} is the largest-magnitude zero of the polynomial $z^{m-1} - z^{m-2} - \cdots - 1$. Moreover, we completely characterize the sequences A and B that achieve $\max H(A,B)$, as well as those that achieve $R(2,m,n)$ for all sufficiently large n.

Key words. Shannon capacity, constrained codes, periodic prefix-synchronized codes, shifts of finite type

AMS subject classifications. 68P30, 94A45, 94A55, 37B10

DOI. 10.1137/S0895480102402757

1. Introduction.

We begin by defining the notion of Shannon capacity [1], [13] of a constrained system of binary sequences. Given a constraint set (or forbidden set) $\mathcal{F}$ of finite-length binary sequences, we define the corresponding *constrained system*, $\mathcal{S}(\mathcal{F})$, to be the set of finite-length binary sequences that do not contain any member of $\mathcal{F}$ as a contiguous subsequence. The *Shannon capacity* of the constrained system $\mathcal{S}(\mathcal{F})$ is defined as $H(\mathcal{F}) = \lim_{n\to\infty} n^{-1} \log_2 q_{\mathcal{F}}(n)$, where $q_{\mathcal{F}}(n)$ is the number of length-n sequences in $\mathcal{S}(\mathcal{F})$. In this paper, we shall be concerned with constraint sets $\mathcal{F} \subset \{0,1\}^m$ containing binary sequences of a fixed length m and, for the most part, sets $\mathcal{F}$ of cardinality 2. The main contribution of this paper is a complete solution to the problem of finding the set $\{A,B\}$ that maximizes $H(A,B)$[1] among all 2-subsets of $\{0,1\}^m$.

The motivation for this problem comes from two sources. The first source is the area of symbolic dynamics [10], where the problem may be reformulated in terms of characterizing those shifts of finite type that have the maximum entropy among all shifts that forbid 2-subsets of $\{0,1\}^m$. Indeed, a related problem was considered by Lind [11], who provided computable bounds on the change in the entropy of a shift of finite type when an extra sequence is added to the original set of forbidden sequences.

The second source for the problem is a closely related question that arises in the context of periodic prefix-synchronized (PPS) codes, which were introduced recently [8] as a family of sync-timing codes. Sync-timing codes are needed in most communication systems where data synchronization is needed (cf. [12]), i.e., when a

[*] Received by the editors February 19, 2002; accepted for publication (in revised form) June 17, 2003; published electronically November 14, 2003. This work was supported by NSF grants NCR-9415754 and CCR-9815006.

http://www.siam.org/journals/sidma/17-2/40275.html.

[†] Department of Electrical and Computer Engineering, University of California, San Diego, CA 92093 (nkashyap@ece.ucsd.edu).

[1] For ease of notation, we use $H(A)$, $H(A,B)$, etc. instead of $H(\{A\})$, $H(\{A,B\})$, etc.

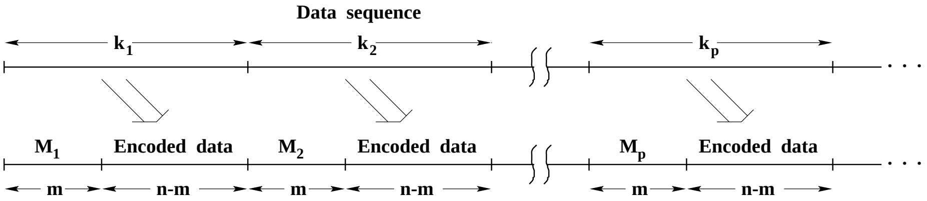


FIG. 1. *A fragment of the encoded bitstream.*

sequence of data symbols must be encoded into bits and transmitted across a channel that can make arbitrary insertion, deletion, and substitution errors. These codes not only enable the decoder to resynchronize rapidly upon cessation of such errors to correctly reproduce data symbols, but also allow the decoder to produce estimates of the time indices of the decoded data symbols in order to determine their positions in the original source sequence.

PPS codes are binary block codes that insert synchronizing markers at regular intervals (see Figure 1). They are characterized by positive integers p, m, n, distinct binary length-m sequences called markers $M_1, \dots, M_p$, and codebooks $C_1, \dots, C_p$. Each C_i contains all binary codewords of length $n > m$, with the following properties: (i) each codeword begins with the marker M_i, and (ii) if $M_i = a_1^{(i)} \dots a_m^{(i)}$, $i = 1, 2, \dots, p$, and $a_1^{(i)} \dots a_m^{(i)} b_1 \dots b_{n-m}$ is a codeword from C_i, then none of the markers $M_1, \dots, M_p$ can be found as a contiguous subsequence of $a_2^{(i)} \dots a_m^{(i)} b_1 \cdots b_{n-m} a_1^{(i+1)} \dots a_{m-1}^{(i+1)}$ (the superscript $(p+1)$ is to be interpreted as the superscript (1)). The idea here is that if a codeword from C_i were to be followed by one from C_{i+1}, then no marker can appear at any place except at the beginning of each codeword. A specific PPS code with *period* p and markers of length m, whose codebooks contain sequences of length n, will be referred to as a (p, m, n) PPS code.

The sequence of data symbols (which we assume to be binary) to be encoded is first divided into blocks of length $K = \sum_{i=1}^{p} k_i$, where $k_i = \lfloor \log_2 |C_i| \rfloor$, with $|C_i|$ denoting the cardinality of C_i. Each such block is then encoded by the PPS code as follows: the first k_1 data symbols are encoded using the codebook C_1, the next k_2 data symbols are encoded using C_2, and so on until the last k_p data symbols are encoded using C_p. Since the encoding procedure has a block structure with input blocklength K and output blocklength $N = pn$, the rate of the code is $R = K/N$. Note that $R \leq 1$ since $K \leq N$, and it is desirable to have codes with rates as close to 1 as possible, so as to minimize the redundancy introduced.

Due to the constraints on the codewords, it is clear that markers can only appear at specific places in the sequence of encoded bits, and hence any marker can be used by the decoder to recover synchronization. The idea behind inserting multiple markers in a periodic manner in the encoded sequence is that, as explained in [8], this periodicity allows the decoder to estimate the time index of each decoded data symbol, relative to the beginning of the "period" to which it belongs, without compromising on the delay in recovering synchronization.

For $p = 1$, the above description is simply that of a prefix-synchronized code, first studied by Gilbert [3] and further analyzed by Guibas and Odlyzko [4]. The PPS code with markers $M_1 = 000$, $M_2 = 111$ and codebooks $C_1 = \{0001100, 0001010\}$, $C_2 = \{1110011, 1110101\}$ is an example of a $(2, 3, 7)$ PPS code.

Let us define $R(p, m, n)$ to be the maximum rate achievable by a (p, m, n) PPS code, if such a code exists, and to be zero otherwise. To find $R(p,m,n)$, it is necessary to find markers $M_1,\cdots,M_p$ that maximize the sizes of the codebooks $C_1,\cdots,C_p$ subject to the constraints defining the code. Due to the similarity of the constraints involved, it is reasonable to expect that, asymptotically in n, $R(p,m,n)$ is closely related to $H_{p,m} = \max_{\mathcal{F}} H(\mathcal{F})$, with the maximum being taken over all p-subsets $\mathcal{F} \subset \{0,1\}^m$.

The work of Gilbert [3], along with that of Guibas and Odlyzko [4], [5], showed that such a relationship does indeed hold for $p=1$, i.e., for prefix-synchronized codes. Specifically, using generating functions, Gilbert showed that if $m \geq 1$ is fixed, then for all sufficiently large n, $R(1,m,n)$ is achieved by choosing the single marker, M_1, in the code to be either the length-m all-zeros sequence, 0^m, or the length-m all-ones sequence, 1^m. It also follows from his results that $\lim_{n\to\infty} R(1,m,n) = \log_2 \rho_m$, where ρ_m is the largest-magnitude zero of the polynomial $z^m - z^{m-1} - \cdots - 1$. Guibas and Odlyzko subsequently used generating functions to show (among other things) that for any $A \in \{0,1\}^m$, $H(A) \leq H(1^m) = \log_2 \rho_m$ with equality iff $A = 0^m$ or 1^m. Thus, we see that for $m \geq 1$, $\lim_{n\to\infty} R(1,m,n) = H_{1,m} = \log_2 \rho_m$.

In this paper, we derive a corresponding relationship for the case when $p=2$. The major part of this derivation lies in a proof of the fact that for all $m \geq 2$, $H_{2,m} = \log_2 \rho_{m-1}$, where ρ_{m-1} is the largest-magnitude zero of the polynomial $z^{m-1} - z^{m-2} - \cdots - 1$. This fact is then used to show that $\lim_{n\to\infty} R(2,m,n) = \log_2 \rho_{m-1}$ as well. Moreover, we identify (Theorem 1) all the pairs of length-m sequences that achieve $H_{2,m}$, as well as those (Theorem 2) that achieve $R(2,m,n)$ for all sufficiently large n. We also provide a partial result (Theorem 18) for $p=3$, for which we show that $H_{3,m} = \log_2 \rho_{m-1}$ as well, and $H_{3,m}$ is achieved by the set $\{10^{m-1}, 0^{m-1}1, 0^m\}$. However, this set of sequences cannot be used as markers in a $(3,m,n)$ PPS code, as any codeword that begins with 0^m must have an occurrence of 0^m or $0^{m-1}1$ starting at the second bit, which violates the constraints defining the code. Hence, it may not be true that $\lim_{n\to\infty} R(3,m,n) = H_{3,m}$.

For higher values of p, not even $H_{p,m}$ is known, although we conjecture that if $p = 2^k$ for any $k \geq 0$, then for $m \geq k+1$, $H_{p,m} = \log_2 \rho_{m-k}$, where ρ_{m-k} is the largest-magnitude zero of the polynomial $z^{m-k} - \cdots - 1$. This conjecture is based on the following observation. Let $\mathcal{F}_0 = \{0^{m-k}\langle i\rangle_2 : i = 0,1,\ldots,2^{k-1}\}$, where $\langle i\rangle_2$ denotes the k-bit binary representation of i. Also, define the functions $\psi_1,\psi_2,\psi_3 : \{0,1\}^* \to \{0,1\}^*$, where $\{0,1\}^*$ denotes the set of all finite-length binary sequences, as follows: for $b_1b_2\ldots b_n \in \{0,1\}^*$,

$$\begin{aligned}\psi_1(b_1b_2\ldots b_n) &= b_1b_2\ldots b_{n-1},\\ \psi_2(b_1b_2\ldots b_n) &= b_2b_3\ldots b_n,\\ \psi_3(b_1b_2\ldots b_n) &= c_1c_2\ldots c_{n-1},\end{aligned}$$

where $c_i = b_i \oplus b_{i+1}$, $\oplus$ being modulo-2 addition. Note that each ψ_i is a two-to-one function. Now, numerical evidence seems to suggest that when $p = 2^k$ for any p-subset $\mathcal{F} \subset \{0,1\}^m$, $q_{\mathcal{F}}(n) \leq q_{\mathcal{F}_0}(n)$ for all n, with equality (for all n) iff $\mathcal{F} = \psi_{i_1}^{-1} \circ \psi_{i_2}^{-1} \circ \cdots \circ \psi_{i_k}^{-1}(0^{m-k})$ or $\psi_{i_1}^{-1} \circ \psi_{i_2}^{-1} \circ \cdots \circ \psi_{i_k}^{-1}(1^{m-k})$ for any $i_1,i_2,\ldots,i_k \in \{1,2,3\}$. It is a simple matter to verify that when $\mathcal{F}$ is of the above form, we have $q_{\mathcal{F}}(n) = 2^k q_{0^{m-k}}(n-k)$, so that $H(\mathcal{F}) = H(0^{m-k}) = \log_2 \rho_{m-k}$, which leads us to the statement of the conjecture.

Before stating our main result, we define some notation that is used throughout this paper: $\langle 01\rangle_m$ and $\langle 10\rangle_m$ denote the two length-m sequences of alternating 0's and 1's. The main contribution of this paper is a proof of the following result.

THEOREM 1. *If A, B are distinct binary sequences of length $m \geq 5$, then*

$$H(A,B) \leq \log_2 \rho_{m-1}$$

with equality iff $\{A, B\}$ or $\{\overline{A}, \overline{B}\}$ is one of the following: $\{0^m, 1^m\}$, $\{\langle 01\rangle_m, \langle 10\rangle_m\}$, $\{0^m, 10^{m-1}\}$, $\{0^m, 0^{m-1}1\}$, and $\{10^{m-1}, 0^{m-1}1\}$. ($\overline{A}, \overline{B}$ are the sequences obtained by complementing each bit of A, B.)

In the terminology of symbolic dynamics, this theorem shows that the entropy of a shift of finite type that forbids some 2-subset $\mathcal{F} \subset \{0,1\}^m$ is at most $\log_2 \rho_{m-1}$. This maximum is achieved precisely when $\mathcal{F}$ is one of the sets listed in the statement of the theorem.

The approach we use to prove the above theorem is based on a generating function for the number, $q_{AB}(n)$, of length-n binary sequences that do not contain A or B as a contiguous subsequence. This generating function can be expressed in a simple form, based on the concept of correlation between two binary strings, which we now define.

The *correlation* between two binary sequences A and B (not necessarily distinct), denoted by $A \circ B$, is a binary sequence of the same length as A. The ith bit (from the left) of $A \circ B$ is determined as follows: place B under A in such a way that the first bit of B lies under the ith bit of A; if the segments that overlap are identical, then the ith bit of $A \circ B$ is 1, else it is 0. Note that if A and B have the same length, then the first bit of $A \circ B$ is a 1 iff $A = B$. For example, if $A = 110001$ and $B = 1000$, then $A \circ B = 010001$, $B \circ A = 0000$, $A \circ A = 100001$, and $B \circ B = 1000$. The correlation of a sequence A with itself is also called the *autocorrelation* of A.

If $A \circ B = (c_0 c_1 \ldots c_{n-1})$ is the correlation between two sequences A and B, then we define the corresponding *correlation polynomial*

$$\phi_{AB}(z) = \sum_{i=0}^{n-1} c_i z^{n-1-i}. \tag{1}$$

With A and B as in the previous example, we have $\phi_{AB}(z) = z^4 + 1$, $\phi_{BA}(z) = 0$, $\phi_{AA}(z) = z^5 + 1$, and $\phi_{BB}(z) = z^3$. For the sake of notational simplicity, it shall henceforth be tacitly understood that correlation polynomials are functions of the complex variable z and so the argument z will be dropped from their notation whenever deemed necessary.

Guibas and Odlyzko [5] showed that given two distinct sequences $A, B \in \{0,1\}^m$, the generating function for $q_{AB}(n)$, defined by $Q_{AB}(z) = \sum_{n=0}^{\infty} q_{AB}(n) z^{-n}$, can be expressed using correlation polynomials as

$$Q_{AB}(z) = \frac{z(\phi_{AA}\phi_{BB} - \phi_{AB}\phi_{BA})}{(z-2)(\phi_{AA}\phi_{BB} - \phi_{AB}\phi_{BA}) + \phi_{AA} + \phi_{BB} - \phi_{AB} - \phi_{BA}}. \tag{2}$$

Thus, the generating function $Q_{AB}(z)$ is a rational function, and we show that it always has a positive real pole, ρ_{AB}, that is larger in magnitude than any other pole. It then follows from the theory of complex variables[2] (see, e.g., [14, Chap. 5]) that $q_{AB}(n) = c(n)\,(\rho_{AB})^n\,(1 + o(1))$ for some $c(n)$ depending polynomially on n. ($c(n)$ is a constant if ρ_{AB} is a simple pole.) This shows that $H(A,B) = \log_2 \rho_{AB}$, and a careful analysis thereafter shows how ρ_{AB} varies with A and B and what choice of A, B maximizes ρ_{AB}.

[2] We need the largest pole here because we define $Q_{AB}(z)$ as a power series in z^{-1}.

To connect the above theorem with $R(2,m,n)$, we use a rational generating function for the number, $f_{AB}(k)$, of length-k sequences that begin with A, end with B, but do not contain A or B anywhere else. Using Guibas and Odlyzko's methods, it is shown in [9] that if A and B are distinct length-m binary sequences, then for $k > m$, $f_{AB}(k)$ is the coefficient of z^{-k} in the expansion of

$$F_{AB}(z) = \frac{1}{z}\,\frac{(z-2)\phi_{AB}+1}{(z-2)(\phi_{AA}\phi_{BB}-\phi_{AB}\phi_{BA})+\phi_{AA}+\phi_{BB}-\phi_{AB}-\phi_{BA}} \tag{3}$$

as $F_{AB}(z) = \sum_{k=0}^{\infty} v_k z^{-k}$. Note that we can use $f_{AB}(k)$ to define the rate of a $(2,m,n)$ PPS code with markers $M_1 = A$ and $M_2 = B$ as follows:

$$\widehat{R}(A,B,n) = \frac{\lfloor \log_2 f_{AB}(m+n)\rfloor + \lfloor \log_2 f_{BA}(m+n)\rfloor}{2n}. \tag{4}$$

Hence, $R(2,m,n) = \max_{A,B} \widehat{R}(A,B,n)$, the maximum being taken over all pairs of distinct sequences $A, B \in \{0,1\}^m$. By showing that the largest-magnitude pole of $F_{AB}(z)$ is the same, in most cases, as that for $Q_{AB}(z)$, we are able to prove the following result.

THEOREM 2. *If A, B are distinct binary sequences of length $m \geq 5$, then*

$$\lim_{n\to\infty} \widehat{R}(A,B,n) \leq \log_2 \rho_{m-1}$$

with equality iff $\{A,B\} = \{0^m, 1^m\}$ or $\{\langle 01\rangle_m, \langle 10\rangle_m\}$. Thus, $\lim_{n\to\infty} R(2,m,n) = \log_2 \rho_{m-1}$.

This theorem shows that when $m \geq 5$ for all sufficiently large n, $R(2,m,n)$ is either $\widehat{R}(0^m,1^m,n)$ or $\widehat{R}(\langle 01\rangle_m, \langle 10\rangle_m, n)$. In fact, we show further that for nearly all (if not all) values of n, $\widehat{R}(0^m,1^m,n) = \widehat{R}(\langle 01\rangle_m, \langle 10\rangle_m, n)$.

The remainder of this paper is devoted to the proofs of the above results. In section 2, we show that $Q_{AB}(z)$ has a real largest-magnitude pole, ρ_{AB}, by demonstrating that the poles of $Q_{AB}(z)$ are actually eigenvalues of a certain nonnegative matrix, which allows us to utilize the powerful Perron–Frobenius theory. Theorem 1 is proved in section 3 by studying the behavior of ρ_{AB} as A and B vary. In section 4, we explore the relationship between $H(A,B)$ and $\widehat{R}(A,B,n)$ and prove Theorem 2.

2. Walks on graphs. In this section, we show that $Q_{AB}(z)$ has a positive real pole that is largest in magnitude among all poles of $Q_{AB}(z)$. It is well known that $q_{AB}(n)$ is precisely the number of walks of length $n-m+1$ on a certain directed graph $\mathcal{G}_{AB}$ obtained by removing a pair of edges from the de Bruijn graph $\mathcal{G}^{(m-1)}$ of order $m-1$. $\mathcal{G}^{(m-1)}$ is a directed graph with vertex set $\{v_i : i = 0,1,\ldots,2^{m-1}-1\}$ (see Figure 2). If we label each vertex v_i with the $(m-1)$-bit binary representation of i, then $\mathcal{G}^{(m-1)}$ has a directed edge from v_i to v_j iff there exists a binary m-sequence $(b_1, b_2, \ldots, b_m)$ whose first $m-1$ bits form v_i's label and whose last $m-1$ bits form v_j's label. Moreover, this directed edge is labeled with the bit b_m. Thus, each walk of length $n-m+1$ on $\mathcal{G}^{(m-1)}$ has a unique binary n-sequence associated with it, namely the sequence formed by concatenating the label of the initial vertex with the labels of the $n-m+1$ edges constituting the walk. In fact, this establishes a one-to-one correspondence between walks of length $n-m+1$ on $\mathcal{G}^{(m-1)}$ and binary n-sequences. Since the edges of the graph are themselves walks of length 1, there is a one-to-one correspondence between the edge set of $\mathcal{G}^{(m-1)}$ and the set of binary m-sequences. Defining $\mathcal{G}_{AB}$ to be the graph obtained by removing the edges corresponding to the

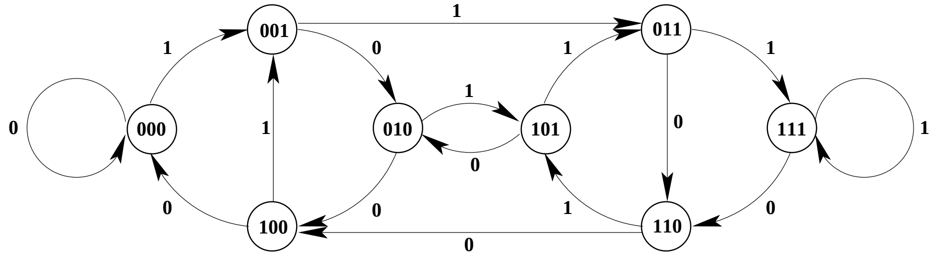


FIG. 2. *The de Bruijn graph* $\mathcal{G}^3$.

sequences A and B from $\mathcal{G}^{(m-1)}$, it is not hard to see that for $n \geq m$, $q_{AB}(n)$ is equal to the number of walks of length $n-m+1$ on $\mathcal{G}_{AB}$. Note that for $0 \leq n \leq m-1$, $q_{AB}(n) = 2^n$, as all the binary sequences of length n do not contain A or B.

Let $\mathcal{A}$ be the adjacency matrix of $\mathcal{G}_{AB}$. It is easy to show that the number of walks of length $n-m+1$ on $\mathcal{G}_{AB}$ is given by the sum of the entries of $\mathcal{A}^{n-m+1}$. Therefore, $q_{AB}(n) = \mathbf{1}^T \mathcal{A}^{n-m+1}\mathbf{1}$ for $n \geq m$, where $\mathbf{1}$ is a column vector of ones. Now if $\mathcal{A} = SJS^{-1}$, where J is the Jordan canonical form of $\mathcal{A}$, then $q_{AB}(n) = \mathbf{1}^T SJ^{n-m+1}S^{-1}\mathbf{1} = \mathbf{x}^T J^{n-m+1}\mathbf{y}$ for some column vectors $\mathbf{x}$ and $\mathbf{y}$. Utilizing the special structure of Jordan forms and working through the details, it can be shown that for $n \geq m$,

$$q_{AB}(n) = \sum_{i=1}^{r} p_i(n)\,(\lambda_i)^n, \tag{5}$$

where $\lambda_1, \ldots, \lambda_r$ are the distinct nonzero eigenvalues of $\mathcal{A}$, and each p_i is a polynomial (possibly zero) of degree strictly less than the algebraic multiplicity of λ_i. Since $\mathcal{A}$ is a nonnegative matrix, the Perron–Frobenius theorem [7, Theorem 8.3.1] shows that it has a real positive eigenvalue, say λ_1, such that $|\lambda_i| \leq \lambda_1$ for $i = 1, \ldots, r$. Note that $\lambda_1 \geq 1$ because if $0 < \lambda_1 < 1$, then (5) shows that we would have $0 < q_{AB}(n) < 1$ for some sufficiently large n. The case $\lambda_1 = 1$ is of little interest, as $q_{AB}(n)$ then grows polynomially with n, which shows that $H(A,B) = 0$. Therefore, we shall henceforth focus on the case when $\lambda_1 > 1$. We next show that in this case, λ_1 is the unique largest-magnitude eigenvalue of $\mathcal{A}$, i.e., $|\lambda_i| < \lambda_1$ for $i \neq 1$, and is algebraically simple. This will then imply that $p_1(n)$ is a nonzero constant.

It is easy to see that since $\mathcal{G}_{AB}$ is obtained by removing two edges from $\mathcal{G}^{(m-1)}$, it is either irreducible (i.e., it has a path connecting every ordered pair of vertices), or it has one vertex with either no incoming edges or no outgoing edges while the rest of the graph is irreducible. It is also a straightforward exercise to show that in the former case, $\mathcal{G}_{AB}$ is aperiodic as well (i.e., the greatest common divisor of the lengths of all the cycles in the graph is 1), because it either contains a loop (an edge that connects a vertex to itself), or it contains two cycles of lengths 2 and 3, respectively. Therefore, by Theorem 4.5.11 in [10], $\mathcal{A}$ has a unique largest-magnitude eigenvalue which is also algebraically simple. In the case when $\mathcal{G}_{AB}$ has an isolated vertex, if we let $\widehat{\mathcal{G}}_{AB}$ be the subgraph of $\mathcal{G}_{AB}$ obtained by removing that vertex and all the edges attached to it, then $\widehat{\mathcal{G}}_{AB}$ is also aperiodic as it always contains a loop. If $\hat{\mathcal{A}}$ is the adjacency matrix of $\widehat{\mathcal{G}}_{AB}$, then $\hat{\mathcal{A}}$ has a unique largest eigenvalue which is also algebraically simple. Now, the eigenvalues of $\mathcal{A}$ are precisely the eigenvalues of $\hat{\mathcal{A}}$, along with the eigenvalues of the adjacency matrix of the subgraph of $\mathcal{G}_{AB}$ formed by the isolated vertex and any edges connecting that vertex to itself [10, section 4.4].

Since the latter adjacency matrix contains only one element, which is either 0 or 1, the only eigenvalue it contributes to $\mathcal{A}$ is 0 or 1. Hence, as $\lambda_1 > 1$, it must come from $\hat{\mathcal{A}}$ and so it is the unique largest-magnitude eigenvalue of $\mathcal{A}$ and is algebraically simple.

Since $p_1(n)$ has degree strictly less than the algebraic multiplicity of λ_1, we see that $p_1(n)$ is a constant c_1. Now, it is known that the number of length-n walks on any graph is at least as large as $c\,(\lambda_{\max})^n$, where $\lambda_{\max}$ is the largest eigenvalue of the adjacency matrix of the graph, and c is some positive constant. Therefore, we have $q_{AB}(n) \geq c(\lambda_1)^n$. Since $\lambda_1 > 1$ is the unique largest-magnitude eigenvalue of $\mathcal{A}$, (5) now shows that $p_1(n) = c_1 > 0$. (At this point, we would like to remark that even if $\lambda_1 = 1$, then it can be argued that $p_1(n)$ is nonzero but not necessarily a constant, but we omit the argument as this fact is not essential for our results.)

Let us now define the function $\widehat{Q}_{AB}(z) = \sum_{n=m}^{\infty} q_{AB}(n) z^{-n}$. Since $q_{AB}(n)$, $n \geq m$, can be expressed in the form given in (5), using the formulae for summation of infinite series, it is easy to see that $\widehat{Q}_{AB}(z)$ is a rational function whose nonzero poles are eigenvalues of $\mathcal{A}$, and the multiplicity of the pole at λ_i is precisely one larger than the degree of p_i. (If $p_i \equiv 0$, then we take $\deg(p_i)$ to be -1, which implies that there is no pole at λ_i.) Since for $\lambda_1 > 1$, $\deg(p_1) = 0$, there is always a simple pole at λ_1. Note that $Q_{AB}(z) = \sum_{n=0}^{m-1} 2^n z^{-n} + \widehat{Q}_{AB}(z)$, but $\sum_{n=0}^{m-1} 2^n z^{-n}$ cannot contribute any nonzero poles to $Q_{AB}(z)$. Therefore, the nonzero poles of $Q_{AB}(z)$ are the same as those of $\widehat{Q}_{AB}(z)$. In particular, λ_1 is a pole of $Q_{AB}(z)$ and is in fact the unique largest-magnitude pole. (We further remark that in the case when $\lambda_1 = 1$, this argument would show that λ_1 is a largest-magnitude pole of $Q_{AB}(z)$, but it may not be a simple pole).

To summarize, we have shown that $Q_{AB}(z)$ always has a real largest-magnitude pole $\rho_{AB} \geq 1$. If $\rho_{AB} > 1$, then it is the unique largest-magnitude pole and is also simple, and hence as explained previously, $H(A,B) = \log_2 \rho_{AB}$. On the other hand, if $\rho_{AB} = 1$, then $H(A,B) = 0$. Thus, in order to maximize $H(A,B)$, we need to maximize ρ_{AB}. In the next section, we study how ρ_{AB} behaves as A and B vary.

3. Maximizing $\boldsymbol{H(A,B)}$. Note that if we define $D_{AB}(z) = (z-2)(\phi_{AA}\phi_{BB} - \phi_{AB}\phi_{BA}) + \phi_{AA} + \phi_{BB} - \phi_{AB} - \phi_{BA}$, then by adding a zero at 2 to D_{AB}, we obtain a polynomial Δ_{AB} that is easier to handle. More precisely,

$$\Delta_{AB}(z) = (z-2)D_{AB}(z) = \gamma_{AA}\gamma_{BB} - \gamma_{AB}\gamma_{BA}, \tag{6}$$

where γ_{**} is the polynomial defined by $\gamma_{**}(z) = (z-2)\phi_{**}(z) + 1$. Thus, the behavior of $D_{AB}(z)$ is intimately connected with the behavior of the polynomials γ_{**}, making it necessary to gain some understanding of the behavior of γ_{**}.

We shall also find it convenient to define the polynomials $p_k(z) = z^k - z^{k-1} - \cdots - 1$ for $k = 1, 2, \ldots$. It is known [15] that for $k \geq 2$, p_k has exactly one root, which is a simple root, in the region $1 < z < 2$, and all its other roots lie within the unit circle (for $k = 1$, the only root of p_k is 1). We shall denote the largest root of p_k by ρ_k. It is also known that ρ_k increases with k, and $2(1 - 2^{-k}) < \rho_k < 2$ for $k \geq 2$. We use these facts about p_k and ρ_k extensively in all that follows.

To any polynomial ϕ with coefficients 0 and 1, we can associate a *γ-polynomial* defined by $\gamma(z) = (z-2)\phi(z) + 1$. Since $\gamma(z) \geq 1$ for $z \geq 2$, all the real zeros of γ must be less than 2. Moreover, if ϕ is not identically zero, then γ has a real zero in $[1, 2)$ because $\gamma(1) = -\phi(1) + 1 \leq 0$. Thus, the largest positive zero of γ lies in $[1, 2)$. We now provide some more results concerning these γ-polynomials.

Lemma 3. *Let ϕ_1, ϕ_2 be polynomials in z with coefficients* 0 *and* 1*, with $\phi_2(2) = \phi_1(2) + 1$. Let γ_1, γ_2 be the corresponding γ-polynomials. Let k be the largest integer such that the coefficients of z^k in ϕ_1 and ϕ_2 are different. If $k = 0$, then $\gamma_1(z) > \gamma_2(z)$ for all $z < 2$, and if $k \geq 1$, then $\gamma_1(z) \geq \gamma_2(z)$ for $\rho_k \leq z < 2$, with equality iff $z = \rho_k$.*

Proof. Note that the sequence formed by the coefficients of each ϕ_i is the binary representation of the integer $\phi_i(2)$. Therefore, it is convenient to identify each ϕ_i with the binary sequence formed by its coefficients. Since ϕ_1 and ϕ_2 are identified with sequences that are binary representations of successive integers, it can be seen that the coefficient of z^k in ϕ_2 is 1, while that in ϕ_1 is 0. Moreover, for each $j < k$, the coefficient of z^j in ϕ_2 is 0, while that in ϕ_1 is 1. Therefore, $\phi_2(z) - \phi_1(z) = 1$ if $k = 0$, and if $k \geq 1$, $\phi_2(z) - \phi_1(z) = z^k - z^{k-1} - \cdots - 1$.

If $k = 0$, we have $\gamma_2(z) - \gamma_1(z) = (z-2)(\phi_2(z) - \phi_1(z)) = z - 2$, which is negative for $z < 2$. If $k \geq 1$, then $\gamma_2(z) - \gamma_1(z) = (z-2)(\phi_2(z) - \phi_1(z)) = (z-2)\,p_k$. Since ρ_k is the largest root of p_k and $p_k(2) = 1$, we must have $p_k(z) > 0$ for $\rho_k < z \leq 2$, from which the result follows. □

Lemma 4. *Let ϕ_1, ϕ_2 be nonzero polynomials with coefficients* 0 *and* 1*, and let γ_1, γ_2 be the corresponding γ-polynomials. Let r_1 and r_2 be the largest positive roots of γ_1 and γ_2, respectively. Suppose that $\phi_1(2) < \phi_2(2)$. Then, $r_1 \leq r_2$ with equality iff $\phi_1(z) = z^{m-1} + z^{m-2} + \cdots + 1$ and $\phi_2(z) = z^m$ for some $m \geq 1$. Moreover, for $r_2 < z < 2$, $\gamma_1(z) > \gamma_2(z)$.*

Proof. It suffices to consider the case when $\phi_2(2) - \phi_1(2) = 1$. The general case then follows by induction on $\phi_2(2) - \phi_1(2)$. Let k be the largest integer such that the coefficients of z^k in $\phi_1(z)$ and $\phi_2(z)$ differ. Note that $\gamma_1(2) = \gamma_2(2) = 1 > 0$. If $k = 0$, then by the previous lemma, since $r_1 < 2$, we have $\gamma_2(r_1) < \gamma_1(r_1) = 0$, which shows that γ_2 has a real root in $(r_1, 2)$. Therefore, $r_2 > r_1$.

If $k \geq 1$, then define $\hat{\phi}_1(z) = z^{k-1} + \cdots + 1$, $\hat{\phi}_2(z) = z^k$. The corresponding γ-polynomials are $\hat{\gamma}_1(z) = z^k - z^{k-1} - \cdots - 1$, and $\hat{\gamma}_2(z) = (z-1)(z^k - z^{k-1} - \cdots - 1)$. It is clear that if $\phi_i = \hat{\phi}_i$, $i = 1, 2$, then $r_1 = r_2 = \rho_k$.

Now, suppose that $\phi_1 \neq \hat{\phi}_1$, in which case since $\phi_2(2) = \phi_1(2) + 1$, we must have $\phi_2 \neq \hat{\phi}_2$ as well. Note that for $i = 1, 2$, we have $\gamma_i(z) = \hat{\gamma}_i(z) + (z-2)(\phi(z) - \hat{\phi}_i(z))$. Since all the coefficients of z^j, $j \leq k$, in ϕ_i are the same as those in $\hat{\phi}_i$, we see that $\phi_i(z) - \hat{\phi}_i(z)$ is itself a nonzero polynomial with coefficients 0 and 1. Therefore, at $z = \rho_k$, $\hat{\gamma}_i$ vanishes and $\phi_i - \hat{\phi}_i$ is positive, which implies that $\gamma_i(\rho_k) < 0$. Hence, we must have $r_1, r_2 > \rho_k$. Now, again by the previous lemma, we have $g_2(r_1) < g_1(r_1) = 0$, which shows that $r_2 > r_1$.

The fact that $\gamma_1(z) > \gamma_2(z)$ for $r_2 < z < 2$ also follows from the previous lemma, since we have shown that $r_2 \geq \rho_k$. □

Lemma 5. *Let $f(z) = (z-2)p(z)$, where $p(z)$ is a nonzero polynomial of degree m with nonnegative coefficients. Then, $f'(z) > 0$ for $z > 2(1 - \frac{1}{m+1})$.*

Proof. Let $p(z) = \sum_{i=0}^{m} a_i z^i$, with $a_i \geq 0$ for $i = 0, 1, \cdots, m-1$ and $a_m > 0$. Then, $f(z) = \sum_{i=0}^{m} a_i (z-2) z^i$. Therefore, $f'(z) = \sum_{i=0}^{m} a_i ((i+1)z^i - 2iz^{i-1})$. Noting that $(i+1)z^i - 2iz^{i-1} > 0$ for $z > 2(1 - \frac{1}{i+1})$, the lemma follows. □

Lemma 6. *Let ϕ be a polynomial of degree $m \geq 2$ with coefficients* 0 *and* 1*, and let γ be the corresponding γ-polynomial. Then, γ has exactly one real root r in the interval $(1, 2)$. Moreover, r is a simple root, $\gamma(z) < 0$ for $1 < z < r$, and $g(z) > 0$ for $r < z \leq 2$.*

Remark. It is easily verified that the conclusions of the lemma are also valid for $\phi(z) = z + 1$. However, if $\phi(z) = z$, 1, or 0, then γ has no roots in (1,2).

Proof. Define $\hat{\phi}(z) = z^m$ and $\hat{\gamma}(z) = (z-2)z^m + 1 = (z-1)(z^m - z^{m-1} - \cdots - 1)$.

If $\phi = \hat{\phi}$, then $\gamma = \hat{\gamma}$ has exactly one root, ρ_m, in (1,2), and it is simple. Since ρ_m is a simple root, $\hat{\gamma}(z)$ must undergo exactly one sign change in (1,2). Therefore, noting that $\hat{\gamma}(2) = 1 > 0$, we must have $\hat{\gamma}(z) > 0$ for $\rho_m < z \leq 2$, and $\hat{\gamma}(z) < 0$ for $1 < z < \rho_m$.

If $\phi \neq \hat{\phi}$, then $\phi(z) - \hat{\phi}(z) > 0$ for $z > 0$. Therefore, $\gamma(z) - \hat{\gamma}(z) = (z-2)(\phi(z) - \hat{\phi}(z)) < 0$ for $0 < z < 2$. Therefore, $\gamma(z) < \hat{\gamma}(z) < 0$ for $1 < z < \rho_m$, which shows that γ has no roots in $(1, \rho_m)$.

Now by the previous lemma, $\gamma'(z) > 0$ for $z > 2(1 - \frac{1}{m+1})$. Hence, if γ has a root in this range, it must be unique (since γ is strictly increasing), and it must have multiplicity 1 (since $\gamma'(z) \neq 0$). Now by Lemma 4, γ has a root r larger than the largest root, ρ_m, of $\hat{\gamma}(z)$. But as mentioned earlier, $\rho_m > 2(1-2^{-m}) > 2(1-\frac{1}{m+1})$. The negative and positive regions for $\gamma(z)$ are determined using arguments identical to those used above for $\hat{\gamma}(z)$. □

We next prove a theorem that locates the largest positive zero of D_{AB}, which is the denominator of Q_{AB} in (2).

Theorem 7. *Let A and B be distinct binary sequences of length $m \geq 5$. Then, $D_{AB}(z)$ has its largest positive real root ρ in* (1,2). *Moreover,* $\max\{r_{AB}, r_{BA}\} \leq \rho \leq \min\{r_{AA}, r_{BB}\}$*, where r_{**} denotes the largest real root of γ_{**}, and the following statements are all equivalent:*

(a) $\rho = \min\{r_{AA}, r_{BB}\}$;

(b) $\rho = \max\{r_{AB}, r_{BA}\}$;

(c) $\phi_{AA}(z)$ *or* $\phi_{BB}(z) = z^{m-1}$, *and* $\phi_{AB}(z)$ *or* $\phi_{BA}(z) = z^{m-2} + z^{m-3} + \cdots + 1$;

(d) $\{A, B\}$ *or* $\{\overline{A}, \overline{B}\} = \{10^{m-1}, 0^m\}$, $\{10^{m-1}, 0^{m-1}1\}$, *or* $\{0^m, 0^{m-1}1\}$.

($\overline{A}, \overline{B}$ are the sequences obtained by complementing each bit of A, B.)

Remark. If ϕ_{AB} (or ϕ_{BA}) $\equiv 0$, then γ_{AB} (or γ_{BA}) $\equiv 1$, in which case we arbitrarily define r_{AB} (or r_{BA}) to be 0.

Proof. Observe first that since the correlations $A \circ A$ and $B \circ B$ begin with 1, while $A \circ B$ and $B \circ A$ begin with 0, we have $\phi_{AA}(2) - \phi_{AB}(2) \geq 1$ and $\phi_{BB}(2) - \phi_{BA}(2) \geq 1$. Hence, for any $z \geq 2$, we have $D_{AB}(z) \geq 2 > 0$, so that all the real roots of D_{AB} must be less than 2. Recall that $\Delta_{AB}(z) = (z-2)D(z) = \gamma_{AA}\gamma_{BB} - \gamma_{AB}\gamma_{BA}$.

Our first goal is to show $\rho \geq \max\{r_{AB}, r_{BA}\}$. Since $\deg(\phi_{AA}), \deg(\phi_{BB}) \geq 4$, Lemma 6 shows that r_{AA} and r_{BB} are the unique roots of g_{AA} and g_{BB} in $(1,2)$. We first consider the case when $\max\{r_{AB}, r_{BA}\} \in (1,2)$. By Lemma 6, this is the case when $\max\{\deg(\phi_{AB}), \deg(\phi_{BA})\} \geq 2$. It can also be verified that this is the case when either $\phi_{AB}(z)$ or $\phi_{BA}(z)$ is $z+1$. Without loss of generality, suppose $r_{AB} \geq r_{BA}$. Note that by Lemma 4, $r_{AB} \leq r_{AA}$, r_{BB}, and hence by Lemma 6, we must have $g_{AA}(r_{AB}) \leq 0$. A similar argument shows that $g_{BB}(r_{AB}) \leq 0$. Therefore, $\Delta_{AB}(r_{AB}) \geq 0$, which implies that $D_{AB}(r_{AB}) \leq 0$, and since $D_{AB}(2) \geq 2 > 0$, we must have $\rho \geq r_{AB}$.

It remains to consider the case when the possible choices for $\phi_{AB}(z)$ and $\phi_{BA}(z)$ are 0, 1, and z. In all these cases, we have $0 < g_{AB}(1.5)g_{BA}(1.5) \leq 1$. Now,

$$g_{AA}(z) \leq (z-2)z^{m-1} + 1 = -0.5(1.5)^{m-1} + 1$$

for $z = 1.5$. Hence for $m \geq 5$, $g_{AA}(1.5) < -1$, and similarly, $g_{BB}(1.5) < -1$, so that $g_{AA}(1.5)g_{BB}(1.5) > 1$, Therefore, $\Delta_{AB}(1.5) > 0$, which shows that $D_{AB}(1.5) < 0$, and hence $\rho > 1.5$.

We now proceed to show that $\rho \leq \min\{r_{AA}, r_{BB}\}$. Without loss of generality, assume $r_{AA} \leq r_{BB}$. In the region $2 > z > r_{BB} = \max\{r_{AA}, r_{BB}, r_{AB}, r_{BA}\}$, all the γ's are positive. Moreover, Lemma 4 shows that for any z in this region, $\gamma_{AA}(z) <$

$\gamma_{AB}(z)$ and $\gamma_{BB}(z) < \gamma_{BA}(z)$. Therefore, $\Delta_{AB}(z) < 0$ for all $z \in (r_{BB}, 2)$, which means that $\rho \notin (r_{BB}, 2)$.

If $r_{AA} < r_{BB}$ and $r_{AA} < z \leq r_{BB}$, then we must have $\gamma_{AA}(z)$, $\gamma_{AB}(z)$, $\gamma_{BA}(z) > 0$, and $\gamma_{BB}(z) \leq 0$. As a result, $\Delta_{AB}(z) < 0$ in this region, which means that $\rho \notin (r_{AA}, r_{BB}]$. Hence, $\rho \leq r_{AA}$.

It remains to show only that the statements (a), (b), (c), and (d) are all equivalent to one another. We first show that (a) ⇒ (b). Let $\rho = \min\{r_{AA}, r_{BB}\}$, which means that either $\gamma_{AA}(\rho)$ or $\gamma_{BB}(\rho)$ is 0. Therefore, $0 = \Delta_{AB}(\rho) = -\gamma_{AB}(\rho)\gamma_{BA}(\rho)$. Thus, we must have $\rho = r_{AB}$ or r_{BA}. In either case, $\max\{r_{AB}, r_{BA}\} \geq \rho$. The reverse inequality is trivial since $\max\{r_{AB}, r_{BA}\} \leq \min\{r_{AA}, r_{BB}\}$.

(b) ⇒ (a) is proved by a similar argument.

We next show that (a) ⇔ (c). Note first that $\rho = \min\{r_{AA}, r_{BB}\}$ iff r_{AA} or r_{BB} is a root of Δ_{AB}. Since $\Delta_{AB}(r_{AA}) = -\gamma_{AB}(r_{AA})\gamma_{BA}(r_{AA})$, we see that r_{AA} is a root of Δ_{AB} iff r_{AB} or $r_{BA} = r_{AA}$. But by Lemma 4, $r_{AA} = r_{AB}$ or r_{BA} iff $\phi_{AA}(z) = z^{m-1}$ and $\phi_{AB}(z)$ or $\phi_{BA}(z) = z^{m-2} + z^{m-3} + \cdots + 1$. A similar argument shows that r_{BB} is a root of Δ_{AB} iff $\phi_{BB}(z) = z^{m-1}$ and $\phi_{AB}(z)$ or $\phi_{BA}(z) = z^{m-2} + z^{m-3} + \cdots + 1$.

Finally, we show that (c) ⇔ (d). Now, $A \circ B = 01^{m-1}$ (i.e., $\phi_{AB}(z) = z^{m-2} + z^{m-3} + \cdots + 1$) can happen iff the longest proper suffix of A is either 0^{m-1} or 1^{m-1} and is the same as the longest proper prefix of B. Moreover, $A \circ A = 10^{m-1}$ (i.e., $\phi_{AA}(z) = z^{m-1}$) implies that the first and last bits of A are different. Therefore, it easily follows that the correlations listed above can arise iff $\{A, B\}$ or $\{\overline{A}, \overline{B}\}$ is one of the sequence pairs listed in the statement of the proposition. □

Observe that in the above proof, the fact that the polynomials ϕ_{AA}, ϕ_{BB}, ϕ_{AB}, and ϕ_{BA} are correlation polynomials for certain sequences is only used in showing that the statement (d) is equivalent to (c). The rest of the proof continues to work even if we assume only that ϕ_{AA}, ϕ_{BB}, ϕ_{AB}, and ϕ_{BA} are polynomials with coefficients 0 and 1, with $\deg(\phi_{AA}) = \deg(\phi_{BB}) = m - 1 \geq 4$ and $\deg(\phi_{AB}), \deg(\phi_{BA}) < m - 1$. Therefore, the conclusions of the theorem, apart from statement (*d*), remain valid for *any* set of four polynomials ϕ_{AA}, ϕ_{BB}, ϕ_{AB}, and ϕ_{BA} that satisfy the properties listed above. We will, in fact, utilize this observation later.

The following corollary is the first important consequence of the previous theorem.

COROLLARY 8. *For $m \geq 5$, the largest pole (in terms of absolute value) of Q_{AB} in* (2) *is precisely the largest positive real root of D_{AB}.*

Proof. We have already seen earlier that the largest-magnitude pole of Q_{AB} is real and positive. Note that all the poles of Q_{AB} must be roots of D_{AB}. Suppose that the largest positive real root ρ of D_{AB} is not a pole of Q_{AB}. Then, ρ must be a root of the numerator polynomial of Q_{AB}, i.e., we must have $\phi_{AA}(\rho)\phi_{BB}(\rho) = \phi_{AB}(\rho)\phi_{BA}(\rho)$. But then, since $D_{AB}(\rho) = 0$, we also have $\phi_{AA}(\rho) + \phi_{BB}(\rho) = \phi_{AB}(\rho) + \phi_{BA}(\rho)$.

Now, if we have real numbers a, b, c, and d such that $a+b = c+d$ and $ab = cd$, then the polynomials $(z-a)(z-b)$ and $(z-c)(z-d)$ must be identical. This implies that $\{a, b\} = \{c, d\}$. Thus, we have $\{\phi_{AA}(\rho), \phi_{BB}(\rho)\} = \{\phi_{AB}(\rho), \phi_{BA}(\rho)\}$, and hence, $\{\gamma_{AA}(\rho), \gamma_{BB}(\rho)\} = \{\gamma_{AB}(\rho), \gamma_{BA}(\rho)\}$.

By the previous theorem, $\rho \in (1, 2)$ and $\max\{r_{AB}, r_{BA}\} \leq \rho \leq \min\{r_{AA}, r_{BB}\}$. Therefore, by Lemma 6, we must have $\gamma_{AA}(\rho)$, $\gamma_{BB}(\rho) \leq 0$ and $\gamma_{AB}(\rho)$, $\gamma_{BA}(\rho) \geq 0$. Hence, $\{\gamma_{AA}(\rho), \gamma_{BB}(\rho)\} = \{\gamma_{AB}(\rho), \gamma_{BA}(\rho)\}$ iff all of them are 0, i.e., $\rho = r_{AA} = r_{BB} = r_{AB} = r_{BA}$. But by Lemma 4, this is possible iff the correlations $A \circ A$ and $B \circ B$ are 10^{m-1} and $A \circ B$ and $B \circ A$ are 01^{m-1}. However, it is easily seen that no pair of sequences A and B can have this set of correlations, leading to a contradiction that proves the result. □

From now on, we shall denote by ρ_{AB} the largest-magnitude pole of Q_{AB}, which (at least for $m \geq 5$) is also the largest positive root of D_{AB}. Since $\rho_{AB} > 1$ for $m \geq 5$, ρ_{AB} must be a simple pole of $Q_{AB}(z)$, and hence is a simple root of $D_{AB}(z)$. Using the fact that $H(A, B) = \log_2 \rho_{AB}$, we now identify a pair of sequences that *minimizes* $H(A, B)$ among all pairs of binary m-sequences $\{A, B\}$.

Proposition 9. *For $m \geq 5$, $\min\{H(A, B) : A, B \in \{0,1\}^m, A \neq B\}$ is achieved by $A = 110^{m-2}$, $B = 110^{m-4}10$.*

Proof. We shall show that if $\widehat{A}, \widehat{B}$ is any pair of binary m-sequences and $A = 110^{m-2}$, $B = 110^{m-4}10$, then $\rho_{\hat{A}\hat{B}} \geq \rho_{AB}$. Observe first that $\phi_{AA}(z) = \phi_{BB}(z) = z^{m-1}$, and $\phi_{AB} = \phi_{BA} \equiv 0$. Thus, $\Delta_{AB} = {\gamma_{AA}}^2 - 1$. Since $\rho_{AB} \leq r_{AA}$ and r_{AA} is the unique root of γ_{AA} in (1,2), we must have $\gamma_{AA}(\rho_{AB}) \leq 0$.

Now, for any pair of m-sequences $\widehat{A}, \widehat{B}$, since the correlations $\widehat{A} \circ \widehat{A}$ and $\widehat{B} \circ \widehat{B}$ must always begin with 1, we have $\phi_{\hat{A}\hat{A}}(z), \phi_{\hat{B}\hat{B}}(z) \geq z^{m-1} = \phi_{AA}(z)$ for all $z \geq 1$. Therefore, $\gamma_{\hat{A}\hat{A}}(z), \gamma_{\hat{B}\hat{B}}(z) \leq \gamma_{AA}(z)$ for all $z \in [1, 2]$. In particular, we have $\gamma_{\hat{A}\hat{A}}(\rho_{AB}), \gamma_{\hat{B}\hat{B}}(\rho_{AB}) \leq \gamma_{AA}(\rho_{AB}) \leq 0$. Therefore, $\gamma_{\hat{A}\hat{A}}(\rho_{AB})\gamma_{\hat{B}\hat{B}}(\rho_{AB}) \geq (\gamma_{AA}(\rho_{AB}))^2$.

On the other hand, since $\phi_{\hat{A}\hat{B}}, \phi_{\hat{B}\hat{A}} \geq 0$, we see that $\gamma_{\hat{A}\hat{B}}(z), \gamma_{\hat{B}\hat{A}}(z) \leq 1$ for all $z \in [1, 2]$ and, in particular, at $z = \rho_{AB}$. Therefore,

$$\Delta_{\hat{A}\hat{B}}(\rho_{AB}) \geq (\gamma_{AA}(\rho_{AB}))^2 - 1 = \Delta_{AB}(\rho_{AB}) = 0,$$

which shows that $D_{\hat{A}\hat{B}}(\rho_{AB}) \leq 0$. Since $D_{\hat{A}\hat{B}}(2) \geq 2 > 0$, we have $\rho_{\hat{A}\hat{B}} \geq \rho_{AB}$. □

The next lemma, which yields a lower bound to the minimum value of $H(A, B)$, is crucial to the proof of the important theorem that follows it.

Lemma 10. *For $A = 110^{m-2}$, $B = 110^{m-4}10$, $H(A, B) > \log_2 \rho_{m-3}$, where ρ_{m-3} is the largest zero of $z^{m-3} - z^{m-4} - \cdots - 1$.*

Proof. With A, B as above, we have

$$\Delta_{AB}(z) = [(z-2)z^{m-1} + 1]^2 - 1 = (z-2)z^{m-1}[(z-2)z^{m-1} + 2]$$

using $a^2 - b^2 = (a + b)(a - b)$. Therefore, the largest positive zero ρ_{AB} of D_{AB} is the largest positive zero of the polynomial $p(z) = (z - 2)z^{m-1} + 2$. Observe that $p(z) = (z-1)(z^{m-1} - z^{m-2} - \cdots - 1) + 1 = (z-1)[z^2(z^{m-3} - z^{m-4} - \cdots - 1) - z - 1] + 1$. Therefore,

$$\begin{aligned} p(\rho_{m-3}) &= -(\rho_{m-3} - 1)(\rho_{m-3} + 1) + 1 \\ &= 2 - (\rho_{m-3})^2 \leq 2 - (\rho_2)^2 \\ &= 2 - \left(\frac{1 + \sqrt{5}}{2}\right)^2 < 0, \end{aligned}$$

the first inequality arising from the fact that $\rho_{m-3} \geq \rho_2$ for $m \geq 5$. Since $p(2) = 2 > 0$, we must have $\rho_{AB} > \rho_{m-3}$. □

At this point, we introduce a means of comparing two correlation sequences, which will make it easier to comprehend our next result which is a theorem of fundamental importance. Given two correlation sequences $A \circ B = (c_0 c_1 \ldots c_{m-1})$ and $\widehat{A} \circ \widehat{B} = (\hat{c}_0 \hat{c}_1 \ldots \hat{c}_{m-1})$ of the same length, we say that $\widehat{A} \circ \widehat{B}$ is *stronger* than $A \circ B$ and denote it by $\widehat{A} \circ \widehat{B} > A \circ B$ if $\hat{c}_i > c_i$ for the smallest i such that $\hat{c}_i \neq c_i$. (This is simply a lexicographic ordering of the correlation sequences.) Equivalently, $\widehat{A} \circ \widehat{B} > A \circ B$ iff $\phi_{\hat{A}\hat{B}}(2) > \phi_{AB}(2)$. We also define $\widehat{A} \circ \widehat{B} \geq A \circ B$ in the obvious way. We now show that if all the correlations between sequences $\widehat{A}$ and $\widehat{B}$ are stronger than the corresponding correlations between sequences A and B, then $H(\widehat{A},\widehat{B}) \geq H(A,B)$.

Theorem 11. *Let $A, B, \widehat{A}, \widehat{B}$ be binary sequences of length $m \geq 5$ such that $\widehat{A} \circ \widehat{A} \geq A \circ A$, $\widehat{B} \circ \widehat{B} \geq B \circ B$, $\widehat{A} \circ \widehat{B} \geq A \circ B$, and $\widehat{B} \circ \widehat{A} \geq B \circ A$. Then $H(\widehat{A},\widehat{B}) \geq H(A,B)$, with equality iff all the above correlation inequalities hold with equality.*

Instead of directly proving this theorem, we shall find it easier to prove a more general result, which yields the above theorem as a special case. The more general result is easier to state if we introduce the following definition.

Definition. *A quadruple of polynomials $(\phi_1,\phi_2,\phi_3,\phi_4)$, each ϕ_i having coefficients* 0 *and* 1*, is called an* admissible m-quadruple *if* $\deg(\phi_1) = \deg(\phi_2) = m$ *and* $\deg(\phi_3), \deg(\phi_4) < m$.

Observe that if A, B are binary m-sequences, then $(\phi_{AA},\phi_{BB},\phi_{AB},\phi_{BA})$ is an admissible $(m-1)$-quadruple. As observed previously, Theorem 7 is essentially a result on the largest positive root ρ of the polynomial

$$D(z) = (z-2)(\phi_1\phi_2 - \phi_3\phi_4) + \phi_1 + \phi_2 - \phi_3 - \phi_4, \tag{7}$$

where $(\phi_1,\phi_2,\phi_3,\phi_4)$ is an admissible m-quadruple. The proof of that theorem shows that for $m \geq 4$, ρ lies in (1,2) and $\max\{r_3, r_4\} \leq \rho \leq \min\{r_1, r_2\}$, where r_i is the largest positive root of γ_i, the γ-polynomial associated with ϕ_i. Moreover, the equivalence of the corresponding statements (a), (b), and (c) is also established. Furthermore, it clearly follows from the proofs of Proposition 9 and Lemma 10 that for $m \geq 4$, $\rho > \rho_{m-2}$ (ρ_{m-2} being the largest zero of the polynomial $z^{m-2} - \cdots - 1$). These facts will be needed to prove our next result, which covers Theorem 11 as a special case.

Theorem 12. *Let $(\phi_1,\phi_2,\phi_3,\phi_4)$ and $(\hat\phi_1,\hat\phi_2,\hat\phi_3,\hat\phi_4)$ be admissible m-quadruples, $m \geq 4$, such that $\phi_i(2) \leq \hat\phi_i(2)$ for $i = 1,2,3,4$. Let D and $\widehat{D}$ be the corresponding polynomials defined via* (7)*, and let ρ and $\hat\rho$ be their respective largest positive roots. Then $\rho \leq \hat\rho$, with equality iff either $\phi_i(2) = \hat\phi_i(2)$ for $i = 1,2,3,4$, or $\phi_i(z) = \hat\phi_i(z) = z^m$ for $i = 1$ or 2 and $\phi_i(z) = \hat\phi_i(z) = z^{m-1} + z^{m-2} + \cdots + 1$ for $i = 3$ or 4.*

Proof. It should be clear that the following four propositions, when patched together, yield the theorem:

1. If $\phi_1(2) < \hat\phi_1(2)$ and $\phi_i(2) = \hat\phi_i(2)$ for $i = 2,3,4$, then $\rho \leq \hat\rho$ with equality iff $\phi_2(z) = \hat\phi_2(z) = z^m$ and $\phi_i(z) = \hat\phi_i(z) = z^{m-1} + \cdots + 1$ for $i = 3$ or 4.
2. If $\phi_2(2) < \hat\phi_2(2)$ and $\phi_i(2) = \hat\phi_i(2)$ for $i = 1,3,4$, then $\rho \leq \hat\rho$ with equality iff $\phi_1(z) = \hat\phi_1(z) = z^m$ and $\phi_i(z) = \hat\phi_i(z) = z^{m-1} + \cdots + 1$ for $i = 3$ or 4.
3. If $\phi_3(2) < \hat\phi_3(2)$ and $\phi_i(2) = \hat\phi_i(2)$ for $i = 1,2,4$, then $\rho \leq \hat\rho$ with equality iff $\phi_i(z) = \hat\phi_i(z) = z^m$ for $i = 1$ or 2 and $\phi_4(z) = \hat\phi_4(z) = z^{m-1} + \cdots + 1$.
4. If $\phi_4(2) < \hat\phi_4(2)$ and $\phi_i(2) = \hat\phi_i(2)$ for $i = 1,2,3$, then $\rho \leq \hat\rho$ with equality iff $\phi_i(z) = \hat\phi_i(z) = z^m$ for $i = 1$ or 2 and $\phi_3(z) = \hat\phi_3(z) = z^{m-1} + \cdots + 1$.

We shall prove the first proposition alone, as the other propositions can be proved analogously. For the proof of the first proposition, we assume that $\hat\phi_1(2) = \phi_1(2) + 1$ and $\phi_i(2) = \hat\phi_i(2)$ for $i = 2,3,4$. The general case follows by induction on $\hat\phi_1(2) - \phi_1(2)$. As usual, note that $D(2), \widehat{D}(2) \geq 2 > 0$. Let $\gamma_i, \hat\gamma_i$ be the γ-polynomials corresponding to $\phi_i, \hat\phi_i$, $i = 1,2,3,4$, and define the polynomials Δ and $\widehat{\Delta}$ analogous to (6).

We first prove the proposition based on the claim that $\hat\gamma_1(\rho) < \gamma_1(\rho)$, deferring the proof of this claim until later. Note that since ρ cannot exceed the largest positive root, r_2, of γ_2, Lemma 6 shows that $\gamma_2(\rho) \leq 0$ with equality iff $\rho = r_2$. Therefore,

$\hat{\gamma}_1(\rho)\gamma_2(\rho) \geq \gamma_1(\rho)\gamma_2(\rho)$, and hence,

$$\begin{aligned}\widehat{\Delta}(\rho) &= \hat{\gamma}_1(\rho)\hat{\gamma}_2(\rho) - \hat{\gamma}_3(\rho)\hat{\gamma}_4(\rho) = \hat{\gamma}_1(\rho)\gamma_2(\rho) - \gamma_3(\rho)\gamma_4(\rho)\\ &\geq \gamma_1(\rho)\gamma_2(\rho) - \gamma_3(\rho)\gamma_4(\rho) = \Delta(\rho) = (\rho-2)D(\rho) = 0\end{aligned}$$

with equality holding iff $\rho = r_2$. Hence, $\widehat{D}(\rho) \leq 0$ which implies that $\hat{\rho} \geq \rho$, with equality iff $\rho = r_2$. Now, as shown in the proof of Theorem 7, $\rho = r_2$ iff $\phi_2(z) = z^m$ and $\phi_3(z)$ or $\phi_4(z) = z^{m-1} + z^{m-2} + \cdots + 1$.

It remains only to prove the claim that $\hat{\gamma}_1(\rho) < \gamma_1(\rho)$. As observed prior to the statement of the theorem, $\rho_{m-2} < \rho < 2$ for $m \geq 4$. Let k be the largest integer such that the coefficients of z^k in ϕ_1 and $\hat{\phi}_1$ are different. If $k = 0$, then Lemma 3 shows that $\hat{\gamma}_1(\rho) < \gamma_1(\rho)$ as $\rho < 2$. If $1 \leq k \leq m-2$, then Lemma 3 again shows that $\hat{\gamma}_1(\rho) < \gamma_1(\rho)$, since $\rho > \rho_{m-2} \geq \rho_k$.

The case $k = m-1$ arises only when $\phi_1(z) = z^m + z^{m-2} + z^{m-3} + \cdots + 1$ and $\hat{\phi}_1(z) = z^m + z^{m-1}$. Therefore, it will suffice to show that for any admissible m-quadruple $(\phi_1, \phi_2, \phi_3, \phi_4)$ with $\phi_1(z) = z^m + z^{m-2} + z^{m-3} + \cdots + 1$, the corresponding ρ exceeds ρ_{m-1}. Now, an argument similar to the proof of Proposition 9 can be used to show that the ρ corresponding to such an admissible m-quadruple is at least as large as the ρ corresponding to the quadruple $(\phi_1(z), z^m, 0, 0)$, with ϕ_1 as above. Therefore, it is sufficient to show that when ϕ_1 is as above, $\phi_2(z) = z^m$, and $\phi_3 = \phi_4 \equiv 0$, then the largest positive zero of $D(z)$ exceeds ρ_{m-1}. In this case, we have

$$\gamma_2(z) = (z-2)z^m + 1 = (z-1)\left[z(z^{m-1} - z^{m-2} - \cdots - 1) - 1\right],$$

which shows that $\gamma_2(\rho_{m-1}) = (\rho_{m-1} - 1)(-1)$. We also have

$$\begin{aligned}\gamma_1(z) &= (z-2)z^m + (z-2)(z^{m-2} + \cdots + 1) + 1\\ &= \gamma_2(z) - 1 + (z^{m-1} - z^{m-2} - \cdots - 1),\end{aligned}$$

which means that $\gamma_1(\rho_{m-1}) = -\rho_{m-1}$. Therefore,

$$\Delta(\rho_{m-1}) = \rho_{m-1}(\rho_{m-1} - 1) - 1 = (\rho_{m-1})^2 - \rho_{m-1} - 1,$$

which is strictly positive for $m \geq 4$, as $z^2 - z - 1 > 0$ for $z > \rho_2$. This means that $D(\rho_{m-1}) < 0$, and so $\rho > \rho_{m-1}$, thus concluding the proof of the claim and hence the theorem. □

As mentioned previously, Theorem 11 is a special case of Theorem 12. Therefore, the proof of Theorem 11 will be complete if we show that under its hypotheses, we can have a situation where $\widehat{A} \circ \widehat{A}$ or $\widehat{B} \circ \widehat{B} = 10^{m-1}$ and $A \circ B$ or $B \circ A = 01^{m-1}$ only if all the correlation inequalities are satisfied with equality.

Consider the case when $\widehat{A} \circ \widehat{A} = 10^{m-1}$ and $A \circ B = 01^{m-1}$. (We can dispose of the other cases similarly.) For $\widehat{A} \circ \widehat{A} \geq A \circ A$ and $\widehat{A} \circ \widehat{B} \geq A \circ B$ to be true, we must have $A \circ A = 10^{m-1}$ and $\widehat{A} \circ \widehat{B} = 01^{m-1}$ as well. But now, we must have (up to complementation of A, B or $\widehat{A}, \widehat{B}$) $A = 10^{m-1}$, $B = 0^{m-1}b$, $\widehat{A} = 10^{m-1}$, and $\widehat{B} = 0^{m-1}\hat{b}$ for some $b, \hat{b} \in \{0,1\}$. It is easily verified that if $b \neq \hat{b}$, then either $B \circ B > \widehat{B} \circ \widehat{B}$ or $B \circ A > \widehat{B} \circ \widehat{A}$, both of which contradict the hypotheses of the theorem. This completes the proof of Theorem 11.

Having Theorem 11 in hand, we are in a position to begin our search for the binary m-sequences A and B that maximize $H(A, B)$. At this point, it should be noted that if there existed A and B such that $A \circ A = B \circ B = 1^m$ and $A \circ B = B \circ A = 01^{m-1}$, then Theorem 11 would imply that A and B are the sequences for which $H(A, B)$ is a maximum. However, it is a simple exercise to show that no pair of sequences can have these correlations.

In order to prove our next important result, which reduces the search space significantly, we need a couple of preliminary lemmas. Recall that $\langle 01\rangle_m$ and $\langle 10\rangle_m$ denote the two length-m sequences of alternating 0's and 1's.

LEMMA 13. *The only length-m autocorrelation sequence $A \circ A$ stronger than $\langle 10\rangle_m$ is 1^m.*

Proof. If $(b_0 b_1 \dots b_{m-1})$ is an autocorrelation sequence (so that $b_0 = 1$), then it is easily seen that whenever $b_j = 1$ for some $j \in [1, m-1]$, then $b_k = 1$ for any $k \in [1, m-1]$ that is a multiple of j. Moreover, the greatest common divisor (GCD) rule for autocorrelation sequences [6, Theorem 3.1] states that if $b_j = b_k = 1$ for some $j, k \in [1, m-1]$ such that $j + k \leq m + l$, where $l = \gcd(j, k)$, then $b_l = 1$ as well.

Note that any autocorrelation sequence $(b_0 b_1 \dots b_{m-1})$ stronger than $\langle 10\rangle_m$ must have either $b_1 = 1$ or $b_2 = 1$. In the first case, we must have $b_k = 1$ for all $k \leq m-1$. Thus, the only autocorrelation sequence with $b_1 = 1$ is 1^m. On the other hand, if $b_2 = 1$, then $b_i = 1$ for all even i. In addition, if $b_k = 1$ for some odd k, then by the GCD rule applied to the pair of indices $(2, k)$, we must have $b_1 = 1$, and hence $b_k = 1$ for all $k \leq m-1$. Thus, any autocorrelation sequence with $b_2 = 1$ must either be $\langle 10\rangle_m$ or 1^m. □

LEMMA 14. *The only correlation sequence $A \circ B$, between distinct binary m-sequences A and B, that is stronger than $\langle 01\rangle_m$ is 01^{m-1}.*

Proof. Let $A \circ B = (c_0 c_1 \dots c_{m-1})$, with $c_0 = 0$. If $A \circ B > \langle 01\rangle_m$, then c_1 must be 1. Therefore, it follows that $(c_1 \cdots c_{m-1})$ must be the autocorrelation sequence for B', where B' is the sequence obtained from B by deleting its last bit. As a result, for $A \circ B > \langle 01\rangle_m$ to be true, we must have $B' \circ B' > \langle 10\rangle_{m-1}$, which, by the previous lemma, is possible only if $B' \circ B' = 1^{m-1}$. □

We now have the requisite tools to prove a result that considerably simplifies the problem of finding the pair of m-sequences (A, B) that maximizes $H(A, B)$. Recall that $H_{2,m} = \max\{H(A, B) : A, B \in \{0,1\}^m, A \neq B\}$.

PROPOSITION 15. *For $m \geq 5$, if $A, B \notin \{0^m, 1^m\}$, then $H(A, B) \leq \log_2 \rho_{m-1}$ with equality iff $\{A, B\} = \{\langle 01\rangle_m, \langle 10\rangle_m\}$, $\{01^{m-1}, 1^{m-1}0\}$, or $\{10^{m-1}, 0^{m-1}1\}$. Consequently, $H_{2,m} = \max\{H(1^m, B) : B \in \{0,1\}^m, B \neq 1^m\}$.*

Proof. Fix $\widehat{A} = \langle 01\rangle_m$, $\widehat{B} = \langle 10\rangle_m$, so that $\widehat{A} \circ \widehat{A} = \widehat{B} \circ \widehat{B} = \langle 10\rangle_m$ and $\widehat{A} \circ \widehat{B} = \widehat{B} \circ \widehat{A} = \langle 01\rangle_m$. It is easily verified using (2) that

$$Q_{\hat{A}\hat{B}}(z) = \frac{z^{m-1} + z^{m-2} + \cdots + 1}{z^{m-1} - z^{m-2} - \cdots - 1},$$

and hence $H(\widehat{A}, \widehat{B}) = \log_2 \rho_{m-1}$. Therefore, we have $H_{2,m} \geq \log_2 \rho_{m-1}$.

Let $A, B \notin \{0^m, 1^m\}$ be a pair of distinct binary m-sequences. Then, we either have $\widehat{A} \circ \widehat{A} \geq A \circ A$, $\widehat{B} \circ \widehat{B} \geq B \circ B$, $\widehat{A} \circ \widehat{B} \geq A \circ B$, and $\widehat{B} \circ \widehat{A} \geq B \circ A$, or at least one of these correlation inequalities is not satisfied. In the former case, Theorem 11 shows that $H(\widehat{A}, \widehat{B}) \geq H(A, B)$ with equality iff $A \circ A = B \circ B = \langle 10\rangle_m$ and $A \circ B = B \circ A = \langle 01\rangle_m$, which can happen iff $\{A, B\} = \{\langle 01\rangle_m, \langle 10\rangle_m\}$.

We next deal with the case when $A \circ B > \widehat{A} \circ \widehat{B}$ which, by Lemma 14, means that $A \circ B = 01^{m-1}$. Therefore, up to complementation of A and B, we must have $\{A, B\} = \{01^{m-1}, 1^{m-1}0\}$, $\{01^{m-1}, 1^m\}$, or $\{1^m, 1^{m-1}0\}$. Our assumption that $A, B \notin \{0^m, 1^m\}$ eliminates the last two sequence pairs along with their complements. When $\{A, B\} = \{01^{m-1}, 1^{m-1}0\}$, it can be verified that $D_{AB}(z) = (z -$

$1)z^{m-1}(z^{m-1}-z^{m-2}-\cdots-1)$, so that $H(A,B)=\log_2\rho_{m-1}=H(\widehat{A},\widehat{B})$. This shows that when $A\circ B>\widehat{A}\circ\widehat{B}$, we must have $H(A,B)=H(\widehat{A},\widehat{B})=\log_2\rho_{m-1}$. The case when $B\circ A>\widehat{B}\circ\widehat{A}$ is similar and leads to the same conclusion.

We are left with the case when we have $A\circ A>\widehat{A}\circ\widehat{A}$ or $B\circ B>\widehat{B}\circ\widehat{B}$, so that by Lemma 13, either $A\circ A$ or $B\circ B$ is 1^m, i.e., A or $B=1^m$ or 0^m. But this is not possible, as we assumed that $A,B\notin\{0^m,1^m\}$. Therefore, one of the previously considered cases must hold, and hence $H(A,B)\le\log_2\rho_{m-1}$ with equality iff $\{A,B\}$ is one of the pairs listed in the statement of the proposition.

Finally, it is easily verified that $Q_{0^m1^m}(z)=Q_{\hat{A}\hat{B}}(z)$, and hence $H(0^m,1^m)=H(\widehat{A},\widehat{B})=\log_2\rho_{m-1}$. Therefore, if the inequality $H_{2,m}\ge\log_2\rho_{m-1}$ is actually an equality, then one of the maximizing pairs $\{A,B\}$ contains 1^m. On the other hand, if this inequality is strict, then any of the sequence pairs that achieve the maximum must include either 1^m or 0^m. But since $H(A,B)=H(\overline{A},\overline{B})$, where $\overline{A}$ and $\overline{B}$ are the sequences obtained by complementing each bit of A and B, at least one of the maximizing pairs includes 1^m, which concludes the proof of the proposition. $\square$

We have thus reduced the problem of maximizing $H(A,B)$ to the problem of finding the sequence $B\neq1^m$ that maximizes $H(1^m,B)$. We tackle this problem by considering the following two cases separately: (i) B begins or ends with a 0, and (ii) B begins and ends with a 1. In fact, as explained below, it is possible to reduce the search space even further in each of these cases.

Given a sequence $A=(a_1a_2\dots a_{n-1}a_n)$, let A^R denote the sequence obtained by reversing A, i.e., $A^R=(a_na_{n-1}\dots a_2a_1)$. It is clear that if Z is a sequence counted by $q_n(A,B)$ for some A,B, then Z^R is a sequence counted by $q_n(A^R,B^R)$. Therefore, we must have $H(A,B)=H(A^R,B^R)$. This conclusion can also be reached from the observation that $A\circ B=B^R\circ A^R$. In particular, $H(1^m,B)=H(1^m,B^R)$. Thus, if B has a longer run of ones at the end than at the beginning, then the situation is reversed for B^R, but the resulting Shannon capacity is the same in both cases. As a result, for case (i), it suffices to consider only those sequences B that end with a 0, and for case (ii), it is enough to consider sequences that begin with a run of ones that is at least as long as the final run of ones. We deal with case (i) first.

LEMMA 16. *If B is a binary sequence of length $m\ge5$ that begins or ends with a* 0*, then $H(1^m,B)\le\log_2\rho_{m-1}$ with equality iff $B=1^{m-1}0$, 01^{m-1}, or 0^m.*

Proof. It suffices to show that if B ends with a 0, then $H(1^m,B)\le\log_2\rho_{m-1}$ with equality iff $B=1^{m_1}0$, or 0^m. Suppose that $B=1^{m_1}\underline{b}\,0$, where $\underline{b}$ is a binary sequence of length $m-m_1-1$ that begins with a 0, and $0\le m_1\le m-1$. Setting $A=1^m$, we see that $A\circ A=1^m$, $A\circ B=0^{m-m_1}1^{m_1}$, and $B\circ A=0^m$. Moreover, $B\circ B$ must end in a run of m_1 zeros, because that is when, in the procedure used to determine $B\circ B$, one of the initial m_1 1's in B overlaps with the final 0. Thus, we have $\phi_{AA}(z)=z^{m-1}+z^{m-2}+\cdots+1$, $\phi_{AB}(z)=z^{m_1-1}+z^{m_1-2}+\cdots+1$, $\phi_{BA}\equiv0$, and $\phi_{BB}(z)=z^{m-1}+z^{k_1}+z^{k_2}+\cdots+z^{k_r}$ for some $k_1,k_2,\dots,k_r\ge m_1$.

Using the fact that $\gamma_{BA}\equiv1$, we see that

$$\begin{aligned}\Delta_{AB}(z)&=\gamma_{AA}\gamma_{BB}-\gamma_{AB}=\gamma_{AA}\left[(z-2)\phi_{BB}+1\right]-\gamma_{AB}\\&=(z-2)\gamma_{AA}\,\phi_{BB}+\gamma_{AA}-\gamma_{AB}\\&=(z-2)\left[\gamma_{AA}\,\phi_{BB}+(\phi_{AA}-\phi_{AB})\right].\end{aligned}$$

Therefore, $D_{AB}=\gamma_{AA}\,\phi_{BB}+(\phi_{AA}-\phi_{AB})$. Now, $\phi_{AA}(z)-\phi_{AB}(z)=z^{m-1}+z^{m-2}+\cdots+z^{m_1}=\phi_{BB}(z)+\phi(z)$ for some polynomial ϕ with coefficients 0 and 1, since the coefficient of z^k, $0 \leq k \leq m_1 - 1$, in $\phi_{BB}(z)$ is zero. Hence, we can write

$$D_{AB} = (\gamma_{AA} + 1)\phi_{BB} + \phi.$$

Now, $\gamma_{AA}(z) = z^m - z^{m-1} - \cdots - 1 = z(z^{m-1} - z^{m-2} - \cdots - 1) - 1$. Therefore, $\gamma_{AA}(z) \geq -1$ for all $z \geq \rho_{m-1}$, with equality iff $z = \rho_{m-1}$. Hence for any $z \geq \rho_{m-1}$, we have

$$\begin{aligned} D_{AB}(z) &\geq \phi(z) && \text{with equality iff } z = \rho_{m-1} \\ &\geq 0 && \text{with equality iff } \phi \equiv 0. \end{aligned}$$

This shows that if $z > \rho_{m-1}$, then $D_{AB}(z) \neq 0$, since the first of the above inequalities is strict. Therefore, $\rho_{AB} \leq \rho_{m-1}$, which shows that $H(A, B) = \log_2 \rho_{AB} \leq \log_2 \rho_{m-1}$.

The above inequalities also show that $D_{AB}(\rho_{m-1}) = 0$ (i.e., $\rho_{AB} = \rho_{m-1}$) iff $\phi \equiv 0$. It is easily verified that with $B = 1^{m-1}0$ or 0^m, we obtain $\phi \equiv 0$. Conversely, if $\phi \equiv 0$, then we must have $\phi_{BB}(z) = z^{m-1} + z^{m-2} + \cdots + z^{m_1}$ or, equivalently, $B \circ B = 1^{m-m_1} 0^{m_1}$. Now, we can either have $m_1 = m - 1$ or $0 \leq m_1 < m - 1$. In the former case, B must be $1^{m-1}0$. In the latter case, since $B \circ B$ begins with two 1's, as shown in the proof of Lemma 13, we must have $B \circ B = 1^m$, which means that B must be 0^m. Therefore, if $\phi \equiv 0$, then B can only be $1^{m-1}0$ or 0^m. This shows that if B ends in a zero, then $H(1^m, B) = \log_2 \rho_{m-1}$ iff $B = 1^{m-1}0$ or 0^m, which concludes the proof of the lemma. □

The only case remaining is when B begins and ends with a 1. We now show that no such B distinct from 1^m can maximize $H(1^m, B)$.

LEMMA 17. *If $B \neq 1^m$ is a binary sequence of length $m \geq 5$ that begins and ends with a 1, then $H(1^m, B) < \log_2 \rho_{m-1}$.*

Proof. As observed prior to the statement of the previous lemma, it is sufficient to consider the case when B is of the form $1^{m_1}\underline{b}1^{m_2}$ with $m_1 \geq m_2 \geq 1$, where $\underline{b}$ is a binary sequence of length $m - m_1 - m_2 > 0$ that begins and ends with a 0. With $A = 1^m$, we have $A \circ A = 1^m$, $A \circ B = 0^{m-m_1}1^{m_1}$ and $B \circ A = 0^{m-m_2}1^{m_2}$. Thus, $\phi_{AA}(z) = z^{m-1} + \cdots + 1$, $\phi_{AB}(z) = z^{m_1-1} + \cdots + 1$, and $\phi_{BA}(z) = z^{m_2-1} + \cdots + 1$. Now, note that since $B \neq 1^m$, $B \circ B$ must begin with 10. Moreover, $B \circ B$ must end with $0^{m_1-m_2}1^{m_2}$, as that is when, in the procedure for determining $B \circ B$, some part of the prefix $1^{m_1}0$ of B overlaps with some part of the suffix 01^{m_2}. Thus, $\phi_{BB}(z) = \phi(z) + \phi_{BA}(z)$, where $\phi(z) = \sum_{k=0}^{m-1} c_k z^k$ is some polynomial with $c_{m-1} = 1$, $c_{m-2} = 0$, $c_k = 0$ for $0 \leq k \leq m_1 - 1$, and the remaining c_k's are either 0 or 1.

With the correlation polynomials being as above, we see that $\gamma_{AA}(z) = z^m - z^{m-1} - \cdots - 1$, $\gamma_{AB}(z) = z^{m_1} - z^{m_1-1} - \cdots - 1$, $\gamma_{BA}(z) = z^{m_2} - z^{m_2-1} - \cdots - 1$, and $\gamma_{BB}(z) = (z-2)\phi + \gamma_{BA}(z)$. Also, note that

$$\begin{aligned} \Delta_{AB}(z) &= \gamma_{AA}[(z-2)\phi + \gamma_{BA}] - \gamma_{AB}\gamma_{BA} \\ &= (z-2)\gamma_{AA}\,\phi + (\gamma_{AA} - \gamma_{AB})\gamma_{BA} \\ &= (z-2)[\gamma_{AA}\,\phi + (\phi_{AA} - \phi_{AB})\gamma_{BA}], \end{aligned}$$

which shows that $D_{AB} = \gamma_{AA}\,\phi + (\phi_{AA} - \phi_{AB})\gamma_{BA}$.

Our goal is to show that $D_{AB}(z) \neq 0$ for all $z \geq \rho_{m-1}$. We claim that D_{AB} is, in fact, an increasing function of z in this region, and so it will suffice to show that $D_{AB}(\rho_{m-1}) > 0$. To justify this claim, we first note that $\gamma_{AA}\,\phi = (z-2)\phi_{AA}\,\phi + \phi$. Since ϕ is a polynomial with coefficients 0 and 1, and $\phi_{AA}\,\phi$ is a polynomial of degree $2m - 2$ with nonnegative coefficients, it follows from Lemma 5 that $\gamma_{AA}\,\phi$ is an increasing function of z for $z > 2(1 - \frac{1}{2m-1})$. Now, $\rho_{m-1} > 2(1 - 2^{-(m-1)}) > 2(1 - \frac{1}{2m-1})$ for $m \geq 5$. Therefore, $\gamma_{AA}\,\phi$ is an increasing function of z for $z \geq \rho_{m-1}$. A similar argument shows that $(\phi_{AA} - \phi_{AB})\gamma_{BA}$ is also increasing in this region, which proves that D_{AB} is an increasing function in the region $z \geq \rho_{m-1}$.

The remainder of the proof just involves finding a positive lower bound for $D_{AB}(\rho_{m-1})$. From now on, for notational simplicity, we shall drop the subscript from ρ_{m-1}. Note first that $(\phi_{AA} - \phi_{AB})(\rho) = \rho^{m-1} + \rho^{m-2} + \cdots + \rho^{m_1} = (\rho^m - \rho^{m_1})/(\rho - 1)$. Next, we have

$$\begin{aligned}\gamma_{BA}(\rho) &= \rho^{m_2} - \rho^{m_2-1} - \cdots - 1\\ &= \rho^{m_2-m+1}(\rho^{m-1} - \rho^{m-2} - \cdots - 1) + \rho^{-1} + \rho^{-2} + \cdots + \rho^{-(m-m_2-1)}\\ &= \frac{\rho^{-1} - \rho^{-(m-m_2)}}{1 - \rho^{-1}} = \frac{1 - \rho^{-(m-m_2-1)}}{\rho - 1}.\end{aligned}$$

Hence, we have

$$\begin{aligned}(\phi_{AA} - \phi_{AB})\,\gamma_{BA}(\rho) &= \frac{1}{(\rho-1)^2}\,(\rho^m - \rho^{m_1} - \rho^{m_2+1} + \rho^{m_1+m_2-(m-1)})\\ &= \frac{1}{(\rho-1)^2}\,[\rho^m - \rho^{m_1} - \rho^{m_1+m_2-(m-1)}(\rho^{m-m_1} - 1)]\\ &\geq \frac{1}{(\rho-1)^2}\,[\rho^m - \rho^{m_1} - (\rho^{m-m_1} - 1)],\end{aligned}$$

the last inequality being a consequence of the fact that $m_1 + m_2 \leq m - 1$, which implies that $\rho^{m_1+m_2-(m-1)} \leq 1$. Noting that $\gamma_{AA}(\rho) = \rho(\rho^{m-1} - \cdots - 1) - 1 = -1$, we obtain

$$D_{AB}(\rho) \geq -\phi(\rho) + \frac{1}{(\rho-1)^2}\,[\rho^m - \rho^{m_1} - (\rho^{m-m_1} - 1)]. \tag{8}$$

We shall first consider the case when $3 \leq m_1 \leq m - 3$. In this case, we see that

$$\begin{aligned}\frac{1}{(\rho-1)^2}\,[\rho^m - \rho^{m_1} - (\rho^{m-m_1} - 1)] &\geq \frac{1}{(\rho-1)^2}\,[\rho^m - \rho^{m-3} - (\rho^{m-3} - 1)]\\ &= \frac{1}{\rho-1}\,[\rho^{m-1} + \rho^{m-2} + \rho^{m-3} - (\rho^{m-4} + \cdots + 1)]\\ &> \frac{1}{\rho-1}\,(\rho^{m-1} + \rho^{m-2}),\end{aligned} \tag{9}$$

the last inequality arising from the fact that $z^{m-3} - z^{m-4} - \cdots - 1 > 0$ for $z > \rho_{m-3}$. Moreover,

$$\phi(\rho) \;=\; \rho^{m-1} + \sum_{k=0}^{m-3} c_k \rho^k \;\leq\; \rho^{m-1} + \sum_{k=0}^{m-3} \rho^k \;<\; \rho^{m-1} + \rho^{m-2} \tag{10}$$

since $z^{m-2} - z^{m-3} - \cdots - 1 > 0$ for $z > \rho_{m-2}$. Putting (9) and (10) into (8), we get

$$\begin{aligned}D_{AB}(\rho) &> -(\rho^{m-1} + \rho^{m-2}) + \frac{1}{\rho-1}\,(\rho^{m-1} + \rho^{m-2})\\ &= \frac{2-\rho}{\rho-1}\,(\rho^{m-1} + \rho^{m-2}) > 0.\end{aligned}$$

This shows that when $3 \le m_1 \le m-3$, $D_{AB}(z) > 0$ for $z = \rho$ and hence for all $z \ge \rho$ as well, which implies that $\rho_{AB} < \rho$ $(= \rho_{m-1})$.

It remains only to deal with the cases when $m_1 = 1$, 2 and $m-2$. Suppose that $m_1 = m-2$. We then have

$$\begin{aligned} \frac{1}{(\rho-1)^2}\left[\rho^m - \rho^{m_1} - (\rho^{m-m_1}-1)\right] &= \frac{1}{(\rho-1)^2}\left[\rho^m - \rho^{m-2} - (\rho^2-1)\right] \\ &= \frac{1}{\rho-1}\left[\rho^{m-1} + \rho^{m-2} - (\rho+1)\right] \\ &\ge \rho^{m-1} + \rho^{m-2} - \rho - 1. \end{aligned} \tag{11}$$

Now, with $m_1 = m-2$, the only possibility for B is $1^{m-2}01$, so that $\phi_{BB}(z) = z^{m-1}+1$, and hence $\phi(z) = z^{m-1}$. Putting this and (11) into (8), we see that $D_{AB}(\rho) \ge \rho^{m-2} - \rho - 1 > \rho^2 - \rho - 1 > 0$.

Next, suppose $m_1 = 2$, which implies that B begins with 110, which in turn means that $B \circ B$ begins with 100. As a result, we have

$$\phi(\rho) \le \rho^{m-1} + \rho^{m-4} + \rho^{m-5} + \cdots + 1 < \rho^{m-1} + \rho^{m-3}, \tag{12}$$

since $z^{m-3} - z^{m-4} - \cdots - 1 > 0$ for $z > \rho_{m-3}$. Note that $\rho^m - \rho^{m_1} - (\rho^{m-m_1}-1)$ is the same for $m_1 = 2$ and $m_1 = m-2$. Therefore, putting (12) and (11) into (8), we get $D_{AB}(\rho) > \rho^{m-2} - \rho^{m-3} - \rho - 1 > 0$, since $z^{m-2} - z^{m-3} - \cdots - 1 > 0$ for $z > \rho_{m-2}$.

Finally, consider $m_1 = 1$, in which case we also have $m_2 = 1$. Therefore, $\phi_{AB} = \phi_{BA} \equiv 1$, and hence $\gamma_{AB}(z) = \gamma_{BA}(z) = z-1$. Therefore,

$$(\phi_{AA} - \phi_{AB})\gamma_{BA}(z) = (z^{m-1} + z^{m-2} + \cdots + z)(z-1) = z^m - z.$$

Also, as in (10), we have $\phi(\rho) < \rho^{m-1} + \rho^{m-2}$. Therefore, since $D_{AB} = \gamma_{AA}\,\phi + (\phi_{AA} - \phi_{AB})\gamma_{BA}$ and $\gamma_{AA}(\rho) = -1$, we see that

$$\begin{aligned} D_{AB}(\rho) &> -(\rho^{m-1} + \rho^{m-2}) + \rho^m - \rho \\ &= \rho(\rho^{m-1} - \rho^{m-2} - \rho^{m-3} - 1), \end{aligned}$$

which is strictly positive for $m \ge 5$, because $z^{m-1} - z^{m-2} - \cdots - 1 = 0$ at $z = \rho = \rho_{m-1}$.

We have thus shown that when B begins and ends with a 1, we have $D_{AB}(z) > 0$ for all $z \ge \rho_{m-1}$, and hence $\rho_{AB} < \rho_{m-1}$, which proves the lemma. □

Putting together the last three results, we obtain Theorem 1.

When $2 \le m \le 4$, it can be verified by computing $H(A,B)$ for all possible A, B of length m that Theorem 1 remains valid when $m = 2$ or 4. When $m = 3$, all but the "only if" part of the theorem remains true. It turns out that for $m = 3$, the maximum Shannon capacity of $\log_2 \rho_2$ is also achieved by two other pairs, namely $\{000, 010\}$ and its complementary pair $\{111, 101\}$.

Interestingly, the answer to the problem of maximizing $H(A,B)$ also provides us with a means of determining the maximum Shannon capacity $H(A,B,C)$ of a constrained system that forbids three distinct binary m-sequences A, B, and C. Formally, let $q_{ABC}(n)$ denote the number of binary n-sequences that do not contain A, B, or C as a contiguous subsequence, and define $H(A,B,C) = \lim_{n\to\infty}(\log_2 q_{ABC}(n))/n$. We now show that $\max_{A,B,C} H(A,B,C) = \max_{A,B} H(A,B) = \log_2 \rho_{m-1}$.

Theorem 18. *For $m \geq 2$,*

$$\max\{H(A,B,C) : A,B,C \in \{0,1\}^m, A \neq B, B \neq C, C \neq A\} = \log_2 \rho_{m-1}.$$

Proof. It is clear that for any three distinct sequences A, B, and C, we have $q_{ABC}(n) \leq q_{AB}(n)$ for all n. Therefore, $H(A,B,C) \leq H(A,B)$, from which it follows that $\max_{A,B,C} H(A,B,C) \leq \max_{A,B} H(A,B)$, where the maximum on the left is taken over all triples of distinct binary m-sequences, and the maximum on the right is taken over all pairs of distinct binary m-sequences.

Now, from Theorem 1 (and, for $2 \leq m \leq 4$, the remarks following the proof of Lemma 17), we know that one of the sequence pairs that achieves $\max_{A,B} H(A,B) = \log_2 \rho_{m-1}$ is $\{10^{m-1}, 0^{m-1}1\}$. Let $\widehat{A} = 10^{m-1}$, $\widehat{B} = 0^{m-1}1$, and $\widehat{C} = 0^m$. For any $\mathcal{F} \subset \{0,1\}^m$, we define $\mathcal{B}_n(\mathcal{F})$ to be the set of all binary n-sequences that do not contain any element of $\mathcal{F}$ as a contiguous subsequence. It is easy to verify that $\mathcal{B}_n(\widehat{A},\widehat{B}) = \mathcal{B}_n(0^{m-1}) \cup \{0^n\}$. Since the only sequence in $\mathcal{B}_n(\widehat{A},\widehat{B})$ that contains 0^m is the all-zeros sequence, it is clear that $\mathcal{B}_n(\widehat{A},\widehat{B},\widehat{C}) = \mathcal{B}_n(0^{m-1})$.

Thus, we see that $q_{\hat{A}\hat{B}\hat{C}}(n) = q_{\hat{A}\hat{B}}(n) - 1$ which shows that $H(\widehat{A},\widehat{B},\widehat{C}) = H(\widehat{A},\widehat{B})$. Since $H(\widehat{A},\widehat{B}) = \log_2 \rho_{m-1}$, we obtain the chain of inequalities

$$\log_2 \rho_{m-1} = H(\widehat{A},\widehat{B},\widehat{C}) \leq \max_{A,B,C} H(A,B,C) \leq \max_{A,B} H(A,B) = \log_2 \rho_{m-1},$$

which proves the theorem. $\square$

4. Connection between $\boldsymbol{H(A,B)}$ and $\boldsymbol{\widehat{R}(A,B,n)}$. We now explore the relationship between the Shannon capacity $H(A,B)$ and the PPS code rate $\widehat{R}(A,B,n)$ defined in (4). We shall show that for nearly all choices of $A,B \in \{0,1\}^m$, $H(A,B) = \lim_{n\to\infty} \widehat{R}(A,B,n)$, and as a result, $\max_{A,B} H(A,B) = \lim_{n\to\infty} R(2,m,n)$, where $R(2,m,n)$ is the maximum possible rate of a $(2,m,n)$ PPS code.

We know from (3) that $F_{AB}(z) = \frac{\gamma_{AB}(z)}{z\,D_{AB}(z)}$ and $F_{BA}(z) = \frac{\gamma_{BA}(z)}{z\,D_{AB}(z)}$ are generating functions for $f_{AB}(k)$ and $f_{BA}(k)$, respectively. Now as noted previously, for $m \geq 5$, the largest positive root, ρ_{AB}, of $D_{AB}(z)$ is simple. Hence, if we establish that ρ_{AB} is also the largest-magnitude pole of $F_{AB}(z)$ and $F_{BA}(z)$, then it would follow that $f_{AB}(k) = c_{AB}\,(\rho_{AB})^k\,(1+o(1))$ and $f_{BA}(k) = c_{BA}\,(\rho_{AB})^k\,(1+o(1))$ for some constants c_{AB} and c_{BA}. This would clearly imply that $\lim_{n\to\infty} \widehat{R}(A,B,n) = \log_2 \rho_{AB} = H(A,B)$. We shall show that ρ_{AB} is almost always the largest-magnitude pole of both $F_{AB}(z)$ and $F_{BA}(z)$, and we shall characterize the exceptional cases.

The first step in this process is to show that ρ_{AB}, which we know is the largest positive root of $D_{AB}(z)$, is in fact the largest-magnitude root of $D_{AB}(z)$. Recall that we have previously shown using Perron–Frobenius theory that whenever $\rho_{AB} > 1$, ρ_{AB} is the unique largest-magnitude pole of $Q_{AB}(z)$ (which is defined by (2)), i.e., if ρ is any other pole of $Q_{AB}(z)$, then $|\rho| < \rho_{AB}$.

Lemma 19. *For $m \geq 5$, if $\rho \neq \rho_{AB}$ is a root of $D_{AB}(z)$, then $|\rho| < \rho_{AB}$.*

Proof. We shall first show that $\rho_{AB} > 1.7$, which, apart from implying that ρ_{AB} is the unique largest pole of $Q_{AB}(z)$, will be important later in the proof. When $m \geq 5$, Proposition 9 shows that $H(A,B)$ is minimized by choosing $A = 110^{m-2}$ and $B = 110^{m-4}10$, and the proof of Lemma 10 shows that for this choice of A and B, $H(A,B) = \log_2 \zeta$, where ζ is the largest real zero of the polynomial $(z-2)z^{m-1} + 2$. Therefore, for any $A,B \in \{0,1\}^m$, $\rho_{AB} \geq \zeta$. For $m \geq 6$, Lemma 10 shows that $\zeta > \rho_3 \approx 1.84$. For $m = 5$, it can be verified that $\zeta \approx 1.816$.

Now, suppose that ρ is a root of $D_{AB}(z)$ such that $|\rho| \geq \rho_{AB}$ and $\rho \neq \rho_{AB}$. We shall first show that ρ must be real (and hence negative) and then reach a contradiction by showing that ρ cannot be less than $-\rho_{AB}$. Since ρ cannot be a pole of $Q_{AB}(z)$, it must be a root of the numerator polynomial of $Q_{AB}(z)$, i.e., $\phi_{AA}(\rho)\phi_{BB}(\rho) = \phi_{AB}(\rho)\phi_{BA}(\rho)$. An argument similar to that in the proof of Corollary 8 now shows that $\{\phi_{AA}(\rho), \phi_{BB}(\rho)\} = \{\phi_{AB}(\rho), \phi_{BA}(\rho)\}$. Thus, ρ must be a root of one of the polynomials $\phi_{AA} - \phi_{AB}$ and $\phi_{AA} - \phi_{BA}$, both of which are polynomials of degree $m-1$ whose coefficients take values in the set $\{0, 1, -1\}$.

A result of Bloch and Pólya [2] states that if $p(z)$ is any polynomial whose coefficients take values in $\{0, 1, -1\}$, then for any $q \in (1, 2)$, the number N of roots of $p(z)$ in the region $|z| > q$ can be bounded as follows:

$$N \leq \frac{1}{2}\left(\log \frac{4q^2}{(3q+1)(q-1)}\right) \Big/ \log\left(1 + \frac{q-1}{2}\right).$$

Evaluating this expression with $q = 1.7$, we see that $p(z)$ has at most one root in the region $|z| > 1.7$.

Thus, since $\rho_{AB} > 1.7$, the polynomials $\phi_{AA} - \phi_{AB}$ and $\phi_{AA} - \phi_{BA}$ can have at most one root in the region $|z| \geq \rho_{AB}$. Since ρ is a root of one of these polynomials, it is the unique root in $|z| > \rho_{AB}$ and hence must be real. Since ρ_{AB} is the largest positive root of D_{AB}, ρ must be negative. Recall from Proposition 9 and Lemma 10 that $\rho_{AB} > \rho_{m-3}$. Thus, we shall reach a contradiction if we can show that no negative root of $\phi_{AA} - \phi_{AB}$ or $\phi_{AA} - \phi_{BA}$ can be less than $-\rho_{m-3}$. We now provide the sketch of an argument that shows this.

Let $(\phi_{AA} - \phi_{AB})(z) = z^{m-1} + \sum_{k=0}^{m-2} c_k z^k$, with the c_k's taking values in $\{0, 1, -1\}$. (The argument for $\phi_{AA} - \phi_{BA}$ is identical.) Suppose first that $m-1$ is even and, further, that c_{m-2} is 0 or -1. In this case, for any $z < -\rho_{m-3}$, we have

$$\begin{aligned}(\phi_{AA} - \phi_{AB})(z) &\geq |z|^{m-1} - \sum_{k=0}^{m-3} |z|^k \\ &= |z|^2 \left(|z|^{m-3} - \sum_{k=0}^{m-4} |z|^k\right) - |z| - 1 + |z|^{m-2} \\ &> |z|^{m-2} - |z| - 1,\end{aligned}$$

with the first inequality holding for any $z < 0$ and the last inequality holding for $|z| > \rho_{m-3}$. But, $|z|^{m-2} - |z| - 1 > 0$ for $m \geq 5$ and $|z| > \rho_2$, which shows that $\phi_{AA} - \phi_{AB}$ has no zeros less than $-\rho_{m-3}$.

Next, suppose that $m-1$ is even and $c_{m-2} = 1$. Then, the correlation $A \circ A$ must begin with 11 and hence must be 1^m. Therefore, $\phi_{AA}(z) = \sum_{k=0}^{m-1} z^k$, which means that $c_k \in \{0, 1\}$ for $k = 0, 1, \ldots, m-3$. We then have for any $z < 0$,

$$\begin{aligned}(\phi_{AA} - \phi_{AB})(z) &\geq |z|^{m-1} - \sum_{\substack{1 \leq k \leq m-2 \\ k \text{ odd}}} |z|^k \\ &= |z|^2 \left(|z|^{m-3} - \sum_{k=0}^{m-4} |z|^k\right) - |z| + \sum_{\substack{2 \leq k \leq m-3 \\ k \text{ even}}} |z|^k \\ &> -|z| + \sum_{\substack{2 \leq k \leq m-3 \\ k \text{ even}}} |z|^k,\end{aligned}$$

with the last inequality holding for $|z| > \rho_{m-3}$. Since $|z|^{m-3} + |z|^{m-5} + \cdots + |z|^2 - |z|$ is clearly positive for $m \geq 5$ and $|z| > 1$, we see that $\phi_{AA} - \phi_{AB}$ has no zeros less than $-\rho_{m-3}$ whenever $m-1$ is even.

A similar argument as above shows that when $m-1$ is odd, then $(\phi_{AA} - \phi_{AB})(z) < 0$ for all $z < -\rho_{m-3}$, which completes the proof of the lemma. ∎

We have thus shown that for $m \geq 5$, ρ_{AB} is the unique largest-magnitude root of D_{AB}. We are now in a position to determine exactly when ρ_{AB} is the largest-magnitude pole of $F_{AB}(z)$ and $F_{BA}(z)$. Note that ρ_{AB} cannot be a pole of *both* $F_{AB}(z)$ and $F_{BA}(z)$ iff ρ_{AB} is a root of γ_{AB} as well as γ_{BA}. But by Theorem 7, this can happen iff $\{A, B\}$ or $\{\overline{A}, \overline{B}\} = \{10^{m-1}, 0^m\}$, $\{10^{m-1}, 0^{m-1}1\}$, or $\{0^m, 0^{m-1}1\}$. This leads us to the following proposition.

PROPOSITION 20. *For all $m \geq 5$, the following are true:*

(a) *If $\{A, B\}$ or $\{\overline{A}, \overline{B}\} = \{0^m, 0^{m-1}1\}$ or $\{0^m, 10^{m-1}\}$, then* $\lim_{n\to\infty} \widehat{R}(A, B, n) = 0$.

(b) *If $\{A, B\}$ or $\{\overline{A}, \overline{B}\} = \{10^{m-1}, 0^{m-1}1\}$, then* $\lim_{n\to\infty} \widehat{R}(A, B, n) = \frac{1}{2} H(A, B)$.

(c) *For all other pairs of distinct binary m-sequences A, B,* $\lim_{n\to\infty} \widehat{R}(A, B, n) = H(A, B)$.

Proof. The discussion preceding the statement of the proposition shows that if $\{A, B\}$ or $\{\overline{A}, \overline{B}\}$ is not one of the pairs listed in (a) and (b), then ρ_{AB} is the unique largest pole, in terms of absolute value, of both $F_{AB}(z)$ and $F_{BA}(z)$. Therefore, as noted prior to the statement of Lemma 19, it follows that $\lim_{n\to\infty} \widehat{R}(A, B, n) = \log_2 \rho_{AB} = H(A, B)$, which proves (c).

To prove (a), note that if $A = 0^m$ and $B = 0^{m-1}1$, then we can have no binary sequence of length $k \geq m+2$ that begins with A and ends with B, but does not contain A or B elsewhere. In other words, $f_{AB}(k) = 0$ for all $k \geq m+2$, and so by definition, $\widehat{R}(A, B, n) = 0$ for all $n \geq m$. The other cases can be similarly dismissed.

Finally, if $A = 10^{m-1}$ and $B = 0^{m-1}1$, then it is clear that the only sequence that can be counted by $f_{AB}(k)$, $k \geq m+2$, is $10^{k-2}1$. Hence, $f_{AB}(k) = 1$ for all $k \geq m+2$. However, $f_{BA}(k) = c_{BA}\,(\rho_{AB})^k\,(1 + o(1))$ for some positive constant c_{BA} because, as can easily be verified, ρ_{AB} is the unique largest-magnitude pole of $F_{BA}(z)$ in this case. As a result, we have $\lim_{n\to\infty} \widehat{R}(A, B, n) = \frac{1}{2} \log_2 \rho_{AB}$, which completes the proof of the proposition. ∎

Theorem 2 is an immediate consequence of the above proposition and Theorem 1. Theorem 2 shows that when $m \geq 5$ for all sufficiently large n, $R(2, m, n)$ is either $\widehat{R}(0^m, 1^m, n)$ or $\widehat{R}(\langle 01\rangle_m, \langle 10\rangle_m, n)$. In fact, as we show next, $|f_{\langle 01\rangle_m\langle 10\rangle_m}(k) - f_{0^m1^m}(k)| \leq 1$ for all k, and hence due to the floor function used in defining $\widehat{R}(A, B, n)$, for nearly all (if not all) values of n, $\widehat{R}(0^m, 1^m, n) = \widehat{R}(\langle 01\rangle_m, \langle 10\rangle_m, n)$. Thus, for nearly all (if not all) sufficiently large integers n,

$$R(2, m, n) = \widehat{R}(0^m, 1^m, n) = \widehat{R}(\langle 01\rangle_m, \langle 10\rangle_m, n).$$

Note that $F_{\langle 01\rangle_m\langle 10\rangle_m}(z) - F_{0^m1^m}(z)$ is a generating function for $f_{\langle 01\rangle_m\langle 10\rangle_m}(k) - f_{0^m1^m}(k)$. Using (3) to get explicit expressions for $F_{\langle 01\rangle_m\langle 10\rangle_m}(z)$ and $F_{0^m1^m}(z)$, we find after some algebraic manipulations that

$$F_{\langle 01\rangle_m\langle 10\rangle_m}(z) - F_{0^m1^m}(z) = \begin{cases} \frac{1}{z(z^m-1)} & \text{if } m \text{ is even,} \\ \frac{z^{m-1}-1}{z^{2m}-1} & \text{if } m \text{ is odd.} \end{cases}$$

It is easily verified that the coefficients in the power series expansions (in the variable z^{-1}) of both $\frac{1}{z(z^m-1)}$ and $\frac{z^{m-1}-1}{z^{2m}-1}$ belong to the set $\{-1, 0, 1\}$. Thus, for each k, $f_{\langle 01\rangle_m\langle 10\rangle_m}(k) - f_{0^m1^m}(k)$ is either -1, 0, or 1.

For the sake of completeness, we would like to mention that when $m = 4$, it can be shown that Theorem 2 remains true in its entirety. When $m = 3$, the theorem remains valid if its statement is modified as follows: $\lim_{n\to\infty} \widehat{R}(A, B, n) \leq \log_2 \rho_2$ with equality iff $\{A, B\} = \{000, 111\}$, $\{010, 101\}$, $\{000, 010\}$, or $\{111, 101\}$. However, it can be shown that if $\{A, B\}$ is one of the last two sequence pairs, then $f_{0^31^3}(k) - f_{AB}(k) = c\,(\rho_2)^k\,(1 + o(1))$, where c is approximately 0.0034. Thus, for all sufficiently large n, we have $\widehat{R}(A, B, n) \leq \widehat{R}(0^m, 1^m, n)$. Finally, when $m = 2$, $\lim_{n\to\infty} \widehat{R}(A, B, n) = 0$ for all sequence pairs A, B.

REFERENCES

[1] J.J. Ashley and P.H. Siegel, *A note on the Shannon capacity of run-length-limited codes*, IEEE Trans. Inform. Theory, 33 (1987), pp. 601–605.

[2] A. Bloch and G. Pólya, *On the roots of certain algebraic equations*, Proc. London Math. Soc. (2), 33 (1930), pp. 102–114; Reproduced in G. Pólya, *Collected Papers: Location of Zeros*, Vol. II, MIT Press, Cambridge, MA, 1974, pp. 336–346.

[3] E.N. Gilbert, *Synchronization of binary messages*, IRE Trans. Inform. Theory, 6 (1960), pp. 470–477.

[4] L.J. Guibas and A.M. Odlyzko, *Maximal prefix-synchronized codes*, SIAM J. Appl. Math., 35 (1978), pp. 401–418.

[5] L.J. Guibas and A.M. Odlyzko, *String overlaps, pattern matching, and nontransitive games*, J. Combin. Theory Ser. A, 30 (1981), pp. 183–208.

[6] L.J. Guibas and A.M. Odlyzko, *Periods in strings*, J. Combin. Theory Ser. A, 30 (1981), pp. 19–42.

[7] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[8] N. Kashyap and D.L. Neuhoff, *Data synchronization with timing*, IEEE Trans. Inform. Theory, 47 (2001), pp. 1444–1460.

[9] N. Kashyap and D.L. Neuhoff, *Periodic prefix-synchronized codes: A generating function approach*, IEEE Trans. Inform. Theory, submitted.

[10] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, Cambridge, UK, 1995.

[11] D.A. Lind, *Perturbation of shifts of finite type*, SIAM J. Discrete Math., 2 (1989), pp. 350–365.

[12] R.A. Scholtz, *Frame synchronization techniques*, IEEE Trans. Comm., 28 (1980), pp. 1204–1212.

[13] C.E. Shannon, *A mathematical theory of communication*, Bell Syst. Tech. J., 27 (1948), pp. 379–423.

[14] H.S. Wilf, *generatingfunctionology*, 2nd ed., Academic Press, San Diego, CA, 1994.

[15] D.A. Wolfram, *Solving generalized Fibonacci recurrences*, Fibonacci Quart., 36 (1998), pp. 129–145.

SIAM J. DISCRETE MATH.
Vol. 17, No. 2, pp. 320–331

ON REGULAR GRAPHS OPTIMALLY LABELED WITH A CONDITION AT DISTANCE TWO[*]

JOHN P. GEORGES[†] AND DAVID W. MAURO[†]

Abstract. For positive integers $j \geq k$, the $\lambda_{j,k}$-number of graph G is the smallest span among all integer labelings of $V(G)$ such that vertices at distance two receive labels which differ by at least k and adjacent vertices receive labels which differ by at least j. We prove that the $\lambda_{j,k}$-number of any r-regular graph is no less than the $\lambda_{j,k}$-number of the infinite r-regular tree $T_\infty(r)$. Defining an r-regular graph G to be (j,k,r)-optimal if and only if $\lambda_{j,k}(G) = \lambda_{j,k}(T_\infty(r))$, we establish the equivalence between (j,k,r)-optimal graphs and r-regular bipartite graphs with a certain edge coloring property for the case $\frac{j}{k} > r$. The structure of r-regular optimal graphs for $\frac{j}{k} \leq r$ is investigated, with special attention to $\frac{j}{k} = 1, 2$. For the latter, we establish that a $(2,1,r)$-optimal graph, through a series of edge transformations, has a canonical form. Finally, we apply our results on optimality to the derivation of the $\lambda_{j,k}$-numbers of prisms.

Key words. $L(j,k)$-labeling, regular graph, prism

AMS subject classification. 05C

DOI. 10.1137/S0895480101391247

1. Introduction.

For positive integers j and k with $j \geq k$, an $L(j,k)$-labeling of graph G is an assignment L of nonnegative integers to the vertices of G such that

(1) $|L(v) - L(u)| \geq j$ if v and u are adjacent, and

(2) $|L(v) - L(u)| \geq k$ if v and u are distance two apart.

Elements of the image of L are called *labels,* and the *span* of L, denoted $s(L)$, is the difference between the largest and smallest labels. The minimum span taken over all $L(j,k)$-labelings of G, denoted $\lambda_{j,k}(G)$, is called the $\lambda_{j,k}$*-number* of G, and if L is a labeling with minimum span, then L is called a $\lambda_{j,k}$*-labeling* of G. Unless otherwise stated, we shall assume with no loss of generality that the minimum label of $L(j,k)$-labelings of G is 0.

A variation of Hale's channel assignment problem [12], the problem of labeling a graph with a condition at distance two, was first investigated in the case $j = 2$ and $k = 1$ by Griggs and Yeh [11]. Other authors have since explored the $\lambda_{2,1}$-numbers of graphs in various classes, as well as relationships between $\lambda_{2,1}(G)$ and other invariants of G (see [2, 6, 9, 10, 13, 14, 16, 17, 18, 19]). Additionally, properties of $\lambda_{j,k}$-numbers have been investigated in [1, 4, 5] and [7].

In this paper, we develop the notion of optimality among r-regular graphs by considering the $\lambda_{j,k}$-number of the infinite r-regular tree $T_\infty(r)$, $r \geq 2$ [4]. We show in section 2 that $\lambda_{j,k}(G) \geq \lambda_{j,k}(T_\infty(r))$ for any r-regular graph G, and we define G to be (j,k,r)*-optimal* if and only if the equality holds. In section 3, we consider the structure of (j,k,r)-optimal graphs for $\frac{j}{k} > r$ and show that (j,k,r)-optimal graphs are bipartite with block edge coloring number r. In section 4, we define the notion of cyclic optimality in the exploration of the case $\frac{j}{k} \leq r$, with special attention to $j = k = 1$. We consider the structure of $(2,1,r)$-optimal graphs in section 5 and

[*] Received by the editors June 15, 2001; accepted for publication (in revised form) May 26, 2003; published electronically November 14, 2003.

http://www.siam.org/journals/sidma/17-2/39124.html

[†] Department of Mathematics, Trinity College, Hartford, CT 06106 (john.georges@trincoll.edu, David.Mauro@trincoll.edu).

establish a canonical form for such graphs. Finally, in section 6, we use the results in the preceding sections to determine the $\lambda_{2,1}$-numbers and $\lambda_{1,1}$-numbers of prisms.

2. Definitions and preliminary results. Throughout the paper, $x \equiv y \pmod{n}$ shall mean that $x - y$ is divisible by n, and $x = y \pmod{n}$ shall mean that x is set equal to the remainder that results when y is divided by n.

Let G be a graph and let L be an $L(j,k)$-labeling of G. Then $M_i(G,L) = \{v \in V(G) \mid L(v) = i\}$ and $m_i(G,L) = |M_i(G,L)|$. When there is no possibility of confusion, reference to G and L will be suppressed.

Georges and Mauro [4] derived $\lambda_{j,k}(T_\infty(r))$ for all j, k, and r, including the two particular cases which will be of importance to this paper.

THEOREM 2.1. *For* $\frac{j}{k} \geq r$, $\lambda_{j,k}(T_\infty(r)) = j + (2r-2)k$.

THEOREM 2.2.

$$\lambda_{j,1}(T_\infty(r)) = \begin{cases} r + 2j - 2 \text{ if } j \leq r, \\ j + 2r - 2 \text{ if } j \geq r. \end{cases}$$

We next show that $\lambda_{j,k}(T_\infty(r))$ is a lower bound for the $\lambda_{j,k}$-numbers of all r-regular graphs, which in turn will serve to motivate the notion of (j,k,r)-optimality.

THEOREM 2.3. *If G is a connected r-regular graph, then* $\lambda_{j,k}(G) \geq \lambda_{j,k}(T_\infty(r))$.

Proof. Suppose L is an $L(j,k)$-labeling of G with span $s(L)$. It suffices to show that L induces an $L(j,k)$-labeling of $T_\infty(r)$ with span $s(L)$.

Let v_{n_0} be an arbitrarily selected vertex in $V(G)$ and let the neighbors of v_{n_0} be $v_{n_1}, v_{n_2}, \ldots, v_{n_r}$. We assign the label $L(v_{n_0})$ to the root w_0 of $T_\infty(r)$, and we assign the labels $L(v_{n_1}), L(v_{n_2}), \ldots, L(v_{n_r})$ to the children $w_1, w_2, \ldots, w_r$ of w_0, respectively. The $r-1$ children of w_i may then be assigned the labels of the neighbors of v_{n_i} which have not already been assigned to the parent of w_i. The result follows by induction. □

For the case $(j,k) = (2,1)$ and $r \geq 2$, the well-known inequality $\lambda_{2,1}(G) \geq r+2$ was used by Jha [13] in his consideration of the λ-number of the Kronecker product of cycles. There, he called those products with $\lambda_{2,1}$-numbers equal to the lower bound *optimal*. We extend his terminology to the consideration of optimal (j,k,r)-labelings of r-regular graphs as follows.

DEFINITION 2.4. *For $r \geq 2$, the graph G is said to be (j,k,r)-optimal if and only if G is r-regular and $\lambda_{j,k}(G) = \lambda_{j,k}(T_\infty(r))$. If L is a $\lambda_{j,k}$-labeling of a (j,k,r)-optimal graph G, then L is said to be a (j,k,r)-optimal labeling of G. We denote the set of (j,k,r)-optimal graphs by $\Gamma(j,k,r)$.*

It follows from Theorems 2.1 and 2.2 that G is $(j,1,r)$-optimal if and only if $\lambda_{j,1}(G) = \lambda_{j,1}(T_\infty(r))$.

3. Optimality with $\frac{j}{k} > r$. In this section we consider the structure of (j,k,r)-optimal graphs for $\frac{j}{k} > r$. As noted in Theorem 2.1, such graphs have $\lambda_{j,k}$-number $j + (2r-2)k$.

THEOREM 3.1. *For $\frac{j}{k} > r$, if G is (j,k,r)-optimal, then G is bipartite with* $|V(G)| \equiv 0 \pmod{2r}$.

Proof. Let L be a (j,k,r)-optimal labeling of G. Since the span of L is $j+(2r-2)k$, each vertex in $V(G)$ has a label in exactly one of the three intervals $X_1 = [0,(r-1)k]$, $X_2 = [(r-1)k+1, j+(r-1)k-1]$, and $X_3 = [j+(r-1)k, j+(2r-2)k]$. Suppose $L(v) \in X_2$, and suppose that exactly m neighbors of v have labels less than $L(v)$, $0 < m < r$. Then the smallest label among the neighbors of v is at most $L(v) - j - (m-1)k$, and the largest label among the neighbors of v is at least $L(v) + j + (r-m-1)k$. The span of L is thus at least $L(v)+j+(r-m-1)k-(L(v)-j-(m-1)k) = 2j+(r-2)k > j+(2r-2)k$, a contradiction. Arguing similarly, if $m = 0$, then the largest label among the neighbors of v is at least $j + (2r-2)k + 1$, a contradiction. And if $m = r$, then the smallest label among the neighbors of v is at most -1, another contradiction. Hence each label assigned by L is in X_1 or X_3. For $i \in \{1, 3\}$, no two distinct vertices in X_i are adjacent since the length X_i is less than j. Hence, G is bipartite.

Now let $v \in V(G)$ with $L(v) \in X_1$. Then the r neighbors of v have labels in X_3. Since the neighbors of v are pairwise distance two apart, their labels pairwise differ by at least k, and hence the labels of the neighbors of v must be $j + (r-1+i)k$, $0 \le i \le r-1$. A similar argument demonstrates that each vertex with label in X_3 has neighbors with labels ik, $0 \le i \le r-1$. Thus, there are exactly $2r$ distinct labels under L with non-zero multiplicity; in X_1, these are $0, k, 2k, \ldots, (r-1)k$, and in X_3 these are $j + (r-1)k, j + rk, \ldots, j + (2r-2)k$.

Let x_1 and x_3 be labels assigned by L in X_1 and X_3, respectively. Then we have seen that each vertex in M_{x_1} is adjacent to some vertex in M_{x_3}. Moreover, due to the distance two condition, no two vertices in M_{x_1} can be adjacent to the same vertex in M_{x_3}. Thus $m_{x_1} \le m_{x_3}$. Similarly, $m_{x_1} \ge m_{x_3}$, implying $m_{x_1} = m_{x_3}$. Hence, since L partitions $V(G)$ into $2r$ nonempty labeling classes, $|V(G)| = 2rm_{x_1}$, from which the result follows. □

We next characterize those graphs in $\Gamma(j, k, r)$, $\frac{j}{k} > r$. It can be easily seen that $K_{r,r}$, the complete r-regular bipartite graph of smallest order, is the graph of smallest order in $\Gamma(j, k, r)$ (see [7]). We also point out that the converse of Theorem 3.1 is not true. For example, the graph $3Q_3$, the sum of 3 copies of the 3-cube, is a 3-regular bipartite graph with order 24; however, for $\frac{j}{k} > 3$, $\lambda_{j,k}(3Q_3) = \lambda_{j,k}(Q_3) = j + 5k$ (see [5]). Alternatively, we observe that a graph G is (j, k, r)-optimal if and only if each component of G is (j, k, r)-optimal. So, since Theorem 3.1 implies that Q_3 is not (j, k, r)-optimal, neither is $3Q_3$.

THEOREM 3.2. *Let G be an r-regular graph with $|V(G)| \equiv 0 \pmod{2r}$. Then $G \in \Gamma(j, k, r)$ if and only if there exists a partition of $V(G)$ into sets $A_0, A_1, A_2, \ldots, A_{r-1}$, $B_0, B_1, B_2, \ldots, B_{r-1}$ such that for each i, $0 \le i \le r-1$, every vertex v in A_i (resp., B_i) has exactly one neighbor in B_j (resp., A_j), $0 \le j \le r-1$.*

Proof. ($\Rightarrow$) Let L be a (j, k, r)-optimal labeling of G. Then the result follows from the proof of Theorem 3.1 with A_i equal to the set of vertices with label ik under L and B_i equal to the set of vertices with label $j + (r-1+i)k$ under L, $0 \le i \le r-1$.

($\Leftarrow$) The vertices in each set A_i (resp., B_i) are pairwise distance three or more apart. Additionally, for $i \ne j$, each vertex in A_i (resp., B_i) is distance two or more from each vertex in A_j (resp., B_i). Thus, we form an $L(j, k)$-labeling L of G by assigning ik to each vertex in A_i, $0 \le i \le r-1$, and $j + (r-1+i)k$ to each vertex in B_i, $0 \le i \le r-1$. Since the span of L is $j + (2r-2)k$, we are done. □

DEFINITION 3.3. *Let $B = X \bigcup Y$ be an r-regular bipartite graph and let $\mathcal{L}$ be an edge coloring of B such that*

(i) *for each $x \in X$, the edges incident to x are assigned the same color under $\mathcal{L}$,*

(ii) *for each $y \in Y$, the edges incident to y are assigned distinct colors under $\mathcal{L}$.*

Then $\mathcal{L}$ is called an X-block coloring of B. We denote the minimum number of colors assigned by X-block colorings of B by $\zeta_X(B)$, and if $\mathcal{L}$ is an X-block coloring of B which assigns exactly $\zeta_X(B)$ distinct colors, then $\mathcal{L}$ is called a minimum X-block coloring of B.

We observe that $r \le \zeta_X(B) \le |X|$. To illustrate, we note that for either bipartition $X \bigcup Y$ of $K_{r,r}$, $\zeta_X(K_{r,r}) = r$ and for any bipartition $X \bigcup Y$ of C_6, $\zeta_X(C_6) = 3$.

Theorem 3.4. *Let G be an r-regular bipartite graph with bipartition W_1 and W_2. Then for $\frac{j}{k} > r$, $G \in \Gamma(j,k,r)$ if and only if $\zeta_{W_1}(G) = \zeta_{W_2}(G) = r$.*

Proof. ($\Rightarrow$) By Theorem 3.2, we let $W_1 = \bigcup_{i=1}^{r-1} A_i$ and $W_2 = \bigcup_{i=1}^{r-1} B_i$. We form a W_1 (resp., W_2)-block coloring using r colors $c_0, c_1, \ldots, c_{r-1}$ by assigning color c_i to each edge which is adjacent to some vertex in A_i (resp., B_i), $0 \le i \le r-1$. But $\zeta_{W_1}(G) \ge r$ (resp., $\zeta_{W_2}(G) \ge r$) since the degree of each vertex in B_0 (resp., A_0) is r. So $\zeta_{W_1}(G) = r$ (resp., $\zeta_{W_2}(G) = r$).

($\Leftarrow$) For $i = 1, 2$, let C_i be minimum W_i-block colorings of G. We produce a vertex labeling L as follows: for each vertex v in W_1 whose incident edges receive color c_i under C_1, $0 \le i \le r-1$, let $L(v) = ik$, and for each vertex in W_2 whose incident edges receive color c_i under C_2, let $L(v) = j + (r-1)k + ik$. To see that L is a (j,k)-labeling, we note that the difference between the largest label among the vertices in W_1 and the smallest label among the vertices in W_2 is $j + (r-1)k - (r-1)k = j$, implying that the labels of adjacent vertices differ by at least j. To show that the distance two condition is satisfied by L, it suffices to show that two vertices distance two apart receive different labels under L. If x_1 and x_2 are distance two apart with $L(x_1) = L(x_2)$ and $x_1, x_2 \in W_1$ (resp., W_2), then there exists vertex $y \in W_2$ (resp., W_1) and edges $\{x_1, y\}$ and $\{x_2, y\}$ which receive the same color under C_1 (resp., C_2), a contradiction. □

We illustrate a 3-regular bipartite graph $B = X \bigcup Y$ on 12 vertices with $\zeta_X(B) = 3$ and $\zeta_Y(B) = 5$. For $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ and $Y = \{y_1, y_2, y_3, y_4, y_5, y_6\}$, let the neighborhood set of x_i, denoted $N(x_i)$, be as follows:

$$\begin{aligned}
N(x_1) &= \{y_1, y_2, y_3\},\\
N(x_2) &= \{y_4, y_5, y_6\},\\
N(x_3) &= \{y_1, y_3, y_5\},\\
N(x_4) &= \{y_2, y_4, y_6\},\\
N(x_5) &= \{y_1, y_3, y_4\},\\
N(x_6) &= \{y_2, y_5, y_6\}.
\end{aligned}$$

Then for $1 \le i \le 3$, assigning color i to the edges incident to x_{2i-1} and x_{2i} shows that $\zeta_X(B) = 3$. On the other hand, examination of the neighborhood sets of each y_i gives $\zeta_Y(B) = 5$.

Theorem 3.5. *For $r \ge 3$, let j, k, j', and k' be integers such that $\frac{j}{k} > r$ and $\frac{j'}{k'} > r$. Then $\Gamma(j,k,r) = \Gamma(j',k',r)$.*

Proof. Let $G \in \Gamma(j,k,r)$. By Theorem 3.1, G is bipartite, so G can be expressed $X \bigcup Y$. This implies $\zeta_X(G) = \zeta_Y(G) = r$ by Theorem 3.4, which in turn implies (also by Theorem 3.4) that $G \in \Gamma(j',k',r)$. Thus $\Gamma(j,k,r) \subseteq \Gamma(j',k',r)$. A similar argument shows $\Gamma(j',k',r) \subseteq \Gamma(j,k,r)$. □

Let $x = \frac{j}{k}$. In [4], it is shown that for any graph G, the function $\lambda_x(G) = \frac{1}{k}\lambda_{j,k}(G)$ is continuous in x on the set of rationals greater than or equal to 1. (Here, continuity at rational number $x \ge 1$ means for any real $\epsilon > 0$, there exists real $\delta > 0$ such that for rational $q \ge 1$ within δ of x, $\lambda_x(G)$ is within ϵ of $\lambda_q(G)$.) Additionally, we have seen that, if $H \in \Gamma(j,k,r)$, $\frac{j}{k} > r$, then $\lambda_x(H) = x + (2r-2)$. Thus $\lambda_r(H) = r + (2r-2)$, which establishes that $H \in \Gamma(ar, a, r)$ for $a \in Z^+$. It follows that for $\frac{j}{k} > r$, $\Gamma(j,k,r) \subseteq \Gamma(ar,a,r)$. We point out, however, that for $\frac{j}{k} > r$, $\Gamma(j,k,r)$ and $\Gamma(ar,a,r)$ are not equal. As an example, K_3, which is not bipartite and hence not optimal for $\frac{j}{k} > 2$, is a member of $\Gamma(2,1,2)$.

4. Optimality with $\frac{j}{k} \in Z^+$ and $\frac{j}{k} < r$. In this section, we investigate the structure of $\Gamma(j,k,r)$ for $\frac{j}{k}$ an integer. Since $\lambda_{j,k}(G) = k\lambda_{c,1}(G)$ for $\frac{j}{k} = c \in Z^+$, it will suffice to assume $k = 1$.

We begin with a consideration of $\Gamma(r-1,1,r)$ for $r \geq 2$.

THEOREM 4.1. *For $r \geq 2$, $\Gamma(r-1,1,r) \subseteq \Gamma(r,1,r)$.*

Proof. If $G \in \Gamma(r-1,1,r)$, then $\lambda_{r-1,1}(G) = 3r-4$ by Theorem 2.2. Let L be a $\lambda_{r-1,1}$-labeling of G. Then $L'(x) = L(x) + \lfloor \frac{L(x)}{r-1} \rfloor$ is an $L(r,1)$-labeling of G with span $3r-2$. □

It follows from the discussion at the end of section 3 that for all $a \in Z^+$ and $\frac{j'}{k'} > r$, $\Gamma(a(r-1),a,r) \bigcup \Gamma(j',k',r) \subseteq \Gamma(ar,a,r)$. We next turn our attention to a special class of optimal labelings.

DEFINITION 4.2. *Let $1 \leq j \leq r$ and let L be a $(j,1,r)$-optimal labeling of r-regular graph G. Then L is said to be a $(j,1,r)$-cyclically optimal labeling of G if and only if for any adjacent vertices v_i and $v_{i'}$ in $V(G)$, $L(v_i) \notin \{L(v_{i'}) \pm j' \pmod{\lambda_{j,1}(G)+1} \mid 0 \leq j' \leq j-1\}$. If G has a $(j,1,r)$-cyclically optimal labeling, then G is said to be $(j,1,r)$-cyclically optimal; otherwise, G is $(j,1,r)$-acyclically optimal. We denote the collection of $(j,1,r)$-cyclically optimal graphs by $\Gamma_c(j,1,r)$.*

To illustrate, we give a $(3,1,4)$-cyclically optimal labeling of a graph in Figure 4.1. We also point out that K_3 is an element of $\Gamma(2,1,2)$ but not of $\Gamma_c(2,1,2)$.

We also note that $\Gamma_c(1,1,r)$ necessarily equals $\Gamma(1,1,r)$.

THEOREM 4.3. *Let G be a $(j,1,r)$-cyclically optimal graph, where $j \leq r$. Then $|V(G)| \equiv 0 \pmod{r+2j-1}$.*

Proof. Let L be a $(j,1,r)$-cyclically optimal labeling of graph G with span $2j+r-2$ by Theorem 2.2. It suffices to show $m_0 = m_1 = \cdots = m_{r+2j-2}$.

By the definition of cyclic labeling, the r neighbors of any vertex v with label $L(v) = x$ must have labels which are precisely the elements of $S_x = \{(L(v)+j+i) \pmod{r+2j-1} \mid 0 \leq i \leq r-1\}$. Thus, since v cannot be adjacent to two vertices with the same label, we have $m_x \leq m_y$ for every y in S_x. But if $y \in S_x$, then $x \in S_y$, so $m_y \leq m_x$. Thus $m_i = m_{j+i} = m_{j+i+1} = m_{i+1}$ for $0 \leq i \leq j+r-3$, giving the result. □

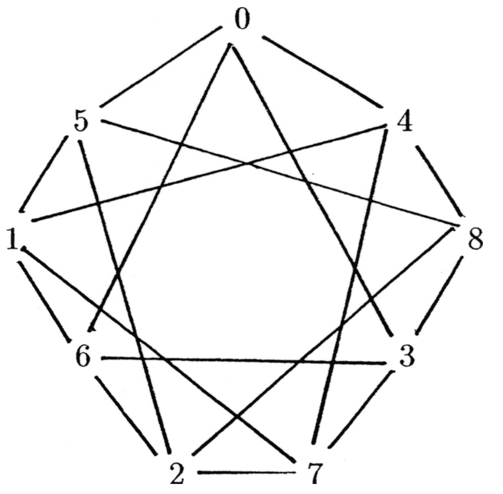


FIG. 4.1. *A $(3,1,4)$-cyclically optimal labeling of graph G.*

COROLLARY 4.4. *If $G \in \Gamma(1,1,r)$, then $|V(G)| \equiv 0 \pmod{r+1}$.*

Proof. If $G \in \Gamma(1,1,r)$, then G is necessarily $(1,1,r)$-cyclically optimal. The result follows immediately from Theorem 3.1. $\square$

We note that the converse to Corollary 4.4 is not true since the $\lambda_{1,1}$-number of $C_4 + C_5$ (the sum of C_4 and C_5) is 4.

THEOREM 4.5. *Let $r \geq 2$. If for fixed j, $1 \leq j \leq r$, G is bipartite and $(j,1,r)$-cyclically optimal, then $|V(G)| \equiv 0 \pmod{2r+4j-2}$.*

Proof. Let L be a $(j,1,r)$-cyclically optimal labeling of the bipartite graph $G = X \bigcup Y$. As in the proof of Theorem 3.1, it can be easily shown that each of the $r+2j-1$ labels has the same multiplicity. Since L is cyclic, the subgraph of G induced by $M_i \bigcup M_{i+j} \bigcup M_{i+2j}$, $0 \leq i \leq r-2$, is 2-regular and thus is a sum of even cycles each of which has order divisible by 6. It follows that $|M_i \bigcap X| = |M_i \bigcap Y|$. Hence, each m_i is even, which establishes the theorem. $\square$

We now give a constructive characterization of $\Gamma_c(j,1,r)$. Let n and h be fixed, $n \geq 3$ and $1 \leq h \leq \lfloor \frac{n}{2} \rfloor$. Then the generalized h-cycle on n vertices, denoted ${}_hC_n$, is the graph with vertex set $\{v_0, v_1, v_2, \ldots, v_{n-1}\}$ and edge set $\{v_i v_s \mid 0 \leq i \leq n-1$ and $s = (i+l) \pmod{n}, 1 \leq l \leq h\}$. We note that ${}_hC_n$ is isomorphic to C_n and K_n when, respectively, $h=1$ and $h = \lfloor \frac{n}{2} \rfloor$.

Now fix r, j, and m, $j \leq r$. For $1 \leq i \leq m$, let G_i be the graph on $r+2j-1$ vertices $v_{i,0}, v_{i,1}, v_{i,2}, \ldots, v_{i,r+2j-2}$ such that for all l, $v_{i,l}$ is adjacent to precisely every vertex in $V(G_i)$ except $v_{i,l \pm x (\mathrm{mod}\, r+2j-1)}$, $0 \leq x \leq j-1$. (We note that G_i is isomorphic to $\overline{{}_hC_n}$, where $h = j-1$ and $n = r+2j-1$.) Then it is easily verified that G_i is in $\Gamma_c(j,1,r)$ and that the labeling L_i of G_i such that $L_i(v_{i,x}) = x$ is a $(j,1,r)$-cyclically optimal labeling. Consequently, the graph $G = \sum_{i=1}^{m} G_i$ is a $(j,1,r)$-cyclically optimal graph and the labeling of G given by $L(v_{i,x}) = x$ is a $(j,1,r)$-cyclically optimal labeling.

Let $\mathcal{M}_0$ be the singleton set containing G, and let $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3 \ldots$ be defined recursively as follows: for $y \geq 1$, $G^{''} \in \mathcal{M}_y$ if and only if for some graph $G^{'} \in \mathcal{M}_{y-1}$ with edges $v_{i_1,x_1} v_{i_1,x_2}$ and $v_{i_3,x_1} v_{i_2,x_2}$, $G^{''}$ results from the following edge transpositions on $G^{'}$:

1. Delete $v_{i_1,x_1} v_{i_1,x_2}$.
2. Delete $v_{i_3,x_1} v_{i_2,x_2}$.
3. Add $v_{i_1,x_1} v_{i_2,x_2}$.
4. Add $v_{i_1,x_2} v_{i_3,x_1}$.

Then by induction, each graph G in $\bigcup_{y=0} \mathcal{M}_y$ is r-regular with cyclic labeling L, since the effect of the edge transpositions is to redirect two edges from vertices labeled x_1 and x_2 to vertices with labels x_1 and x_2. Thus $\bigcup_{y=0} \mathcal{M}_y \subseteq \Gamma_c(j,1,r)$.

To show that $\Gamma_c(j,1,r) \subseteq \bigcup_{y=0} \mathcal{M}_y$, let G be a $(j,1,r)$-cyclically optimal graph and let L be a $(j,1,r)$-cyclically optimal labeling of G. Then Theorem 4.3 implies that $|V(G)| = c(r+2j-1)$ for some c and that we may thus denote the vertices of G by $v_{i,z}$, $1 \leq i \leq c$ and $0 \leq z \leq r+2j-2$, where $L(v_{i,z}) = z$. Furthermore, since the r neighbors of $v_{i,x}$ necessarily have labels $(x+y) \pmod{r+2j-1}$, $j \leq y \leq r+j-1$, it is the case that for every i_1, i_2, x_1, x_2 such that $i_1 \neq i_2$ and v_{i_1,x_1} is adjacent to v_{i_2,x_2}, there exists $i_3 \neq i_1$ such that v_{i_1,x_2} is adjacent to v_{i_3,x_1}. Hence, for $i_1 \neq i_2$, an r-regular graph G' may be formed by executing the following algorithm, which may be thought of as a reversal of the edge manipulation algorithm given above:

a. Delete $v_{i_1,x_1} v_{i_2,x_2}$.
b. Delete $v_{i_1,x_2} v_{i_3,x_1}$.
c. Add $v_{i_1,x_1} v_{i_1,x_2}$.
d. Add $v_{i_3,x_1} v_{i_2,x_2}$.

Moreover, the vertex labeling $L'(v_{i,x})$ of G' given by $L' = L$ is a $(j,1,r)$-optimal labeling, since the effect of these edge manipulations is to redirect two edges from vertices labeled x_1 and x_2 to vertices labeled x_1 and x_2.

As compared to G, the graph G' produced by this algorithm has 1 (or 2) fewer edges of the form $v_{a,x_i}v_{b,x_h}$ where $a \neq b$, and 1 (or 2) more edges of the form $v_{c,x_i}v_{c,x_h}$. The algorithm may thus be iterated sufficiently many times to produce $\sum_{i=1}^{m} G_i$, each of whose edges is of the form $v_{c,x_i}v_{c,x_h}$. Hence, $\Gamma_c(j,1,r) \subseteq \bigcup_{y=0} \mathcal{M}_y$, which in turn implies the following.

Theorem 4.6. *Every $(j,1,r)$-cyclically optimal graph yields, through a sequence of edge transpositions, a graph isomorphic to a sum of copies of $\overline{{}_{j-1}C_{r+2j-1}}$.*

From this construction, we have the following.

Corollary 4.7. *A connected graph G is $(j,1,r)$-cyclically optimal if and only if there exists a partition $\{V_0, V_1, V_2, \ldots, V_{r+2j-2}\}$ of $V(G)$ such that, for $0 \leq i \leq r+2j-2$, each vertex in V_i is adjacent to exactly one vertex in $V_{(i+i')(mod\ r+2j-1)}$, $j \leq i' \leq j+r-1$. Necessarily, the sets in the partition are of equal size.*

5. Optimality with $\frac{j}{k} = 2$. In this section, we investigate the graphs in $\Gamma(2,1,r)$, $r \geq 2$, each of which has $\lambda_{2,1}$-number equal to $r+2$. Since, in general, not all $(2,1,r)$-optimal labelings of r-regular graphs are cyclic, then the $r+3$ labels given by an optimal labeling L need not have equal multiplicities. However, as we shall see, the multiplicities of labels under a $(2,1,r)$-optimal labeling L do possess certain regularities.

We note that for $r = 2$, any graph in $\Gamma(2,1,r)$ is a cycle C_n. Moreover, since $\lambda_{2,1}(C_n) = 4 = r+2$, it follows that $\Gamma(2,1,2) = \{C_n | n \geq 3\}$. It thus suffices to consider $r \geq 3$.

Theorem 5.1. *Let G be a $(2,1,r)$-optimal graph, where $r \geq 3$, and let L be a $\lambda_{2,1}$-labeling of G. Then $m_i = m_h$ for $1 \leq i,h \leq r+1$.*

Proof. Let T be the set of integers in the interval $[0, r+2]$. Then for every integer x, $1 \leq x \leq r+1$, there are exactly r elements in T which differ from x by at least 2. Thus, the distance conditions require the labels of the neighbors of each vertex v with label $L(v) = x$ to be precisely the elements of $S_x = \{w \in T \mid |x - w| \geq 2\}$. So, for every $y \in S_x$, we have $m_x \leq m_y$. But if $y \in S_x$ where $1 \leq y \leq r+1$, then $x \in S_y$, implying $m_y \leq m_x$. Hence $m_h = m_{2+h} = m_{3+h} = m_{h+1}$ for $1 \leq h \leq r-2$, giving the result. □

Now let L be a $(2,1,r)$-optimal labeling of r-regular graph G. We define $M(\alpha,\beta)$ to be the set of vertices in $V(G)$ which have label α and which are adjacent to some vertex with label β, and we denote the cardinality of $M(\alpha,\beta)$ by $m(\alpha,\beta)$. Noting that, for v such that $L(v) = 0$, exactly one element i of the set $\{2,3,4,\ldots,r+2\}$ is not represented among the labels of the neighbors of v, we define $M(0,i^*)$ to be the collection of vertices which are labeled 0 and which are adjacent to no vertices labeled i. For $i \in \{0,1,2,3,4,\ldots,r\}$, we define $M(r+2,i^*)$ analogously, and we denote the cardinalities of $M(0,i^*)$ and $M(r+2,i^*)$ by $m(0,i^*)$ and $m(r+2,i^*)$, respectively. For fixed h, $2 \leq h \leq r+1$, there is a one-to-one correspondence between M_h and $M_0 - M(0,h^*)$, implying that $m_h = m_0 - m(0,h^*)$. So, by Theorem 5.1, it follows that $m(0,2^*) = m(0,3^*) = \cdots = m(0,r+1^*)$. Similarly, for $1 \leq i \leq r$, $m(r+2,1^*) = m(r+2,2^*) = \cdots = m(r+2,r^*)$.

Since $m(0,r+2) = m(r+2,0)$, we have

$$(1) \qquad \sum_{i=2}^{r+1} m(0,i^*) = m(0,r+2) = m(r+2,0) = \sum_{i=1}^{r} m(r+2,i^*).$$

But for every i, $2 \leq i \leq r+1$, each vertex in M_i is adjacent to some vertex in M_0, implying $m(0,i) = m_i$. Thus, since $m(i,0) = m(0,i)$, we have $m_0 = m(0,i) + m(0,i^*) = m_i + m(0,i^*)$, giving $m_0 - m_i = m(0,i^*)$. But $m_2 = m_3 = \cdots = m_{r+1}$ by Theorem 5.1, so $m(0,i^*) = m_0 - m_2$ for every i, $2 \leq i \leq r+1$.

Similarly, $m(r+2,i^*) = m_{r+2} - m_2$, which, by (1), implies

$$\sum_{i=2}^{r+1}(m_0 - m_2) = \sum_{i=1}^{r}(m_{r+2} - m_2).$$

This gives the following theorem.

THEOREM 5.2. *Let $G \in \Gamma(2,1,r)$ and let L be a $\lambda_{2,1}$-labeling of G. Then $m_0 = m_{r+2}$.*

For fixed h, $2 \leq h \leq r+1$, there is a one-to-one correspondence between M_h and $M_0 - M(0,h^*)$, implying that $m_h = m_0 - m(0,h^*)$ (and likewise, $m_h = m_{r+2} - m(r+2,h^*)$ for $1 \leq h \leq r$). So, by Theorems 5.1, 5.2, and (1), it follows that

$$\begin{aligned} &m(0,2^*) = m(0,3^*) = \cdots = m(0,r+1^*)\\ &= m(r+2,1^*) = m(r+2,2^*) = \cdots = m(r+2,r^*). \end{aligned}$$

We use this result to establish the next theorem.

THEOREM 5.3. *Let $G \in \Gamma(2,1,r)$ and let L be a $\lambda_{2,1}$-labeling of G. Then $|V(G)| = (r+3)m(0,r+2^*) + (r^2+2r-1)m(0,2^*)$.*

Proof. By Theorems 5.1 and 5.2, $|V(G)| = \sum_{i=0}^{r+2} m_i = 2m_0 + (r+1)m_2$. Since

$$m_0 = \sum_{i=2}^{r+2} m(0,i^*) = rm(0,2^*) + m(0,r+2^*)$$

and

$$\begin{aligned} m_2 = m_0 - m(0,2^*) &= -m(0,2^*) + \sum_{i=2}^{r+2} m(0,i^*) = \sum_{i=3}^{r+2} m(0,i^*)\\ &= m(0,r+2^*) + \sum_{i=3}^{r+1} m(0,i^*) = m(0,r+2)^* + (r-1)m(0,2^*), \end{aligned}$$

the result now follows via straightforward algebra. □

Since $m(0,2^*)$ and $m(0,r+2^*)$ must be nonnegative, we observe that the smallest $(2,1,r)$-optimal graph has order at least $r+3$. As noted in the preceding section, this bound is achieved by the unique r-regular graph on $r+3$ vertices: $\overline{C_{r+3}}$, which is cyclically optimal. If $G \in \Gamma(2,1,r)$ is acyclically optimal, then every optimal labeling of G has $m(0,r+2) \geq 1$, which in turn implies that $m(0,2^*) \geq 1$. Thus, the smallest $(2,1,r)$-acyclically optimal graph has order at least r^2+2r-1. We produce a $(2,1,r)$-acyclically optimal graph $G_a(r)$ on r^2+2r-1 vertices as follows: noting that $m_0 = m_{r+2} = r$ and $m_1 = m_2 = \cdots = m_{r+1} = r-1$ under an acyclically optimal labeling of G, we define

$$\begin{aligned} &M_i = \{v_i^h, 1 \leq h \leq r-1\} \text{ for } 1 \leq i \leq r+1, \text{ and}\\ &M_i = \{v_i^p, 1 \leq p \leq r\} \text{ for } i = 0, r+2. \end{aligned}$$

Let $P_0 = \{S_1, S_2, S_3, \ldots, S_r\}$, where S_i is lexicographically the ith subset size $r-1$ of $\{2,3,4,\ldots,r+1\}$. Similarly, let $P_{r+2} = \{T_1, T_2, T_3, \ldots, T_r\}$, where T_i is lexico-

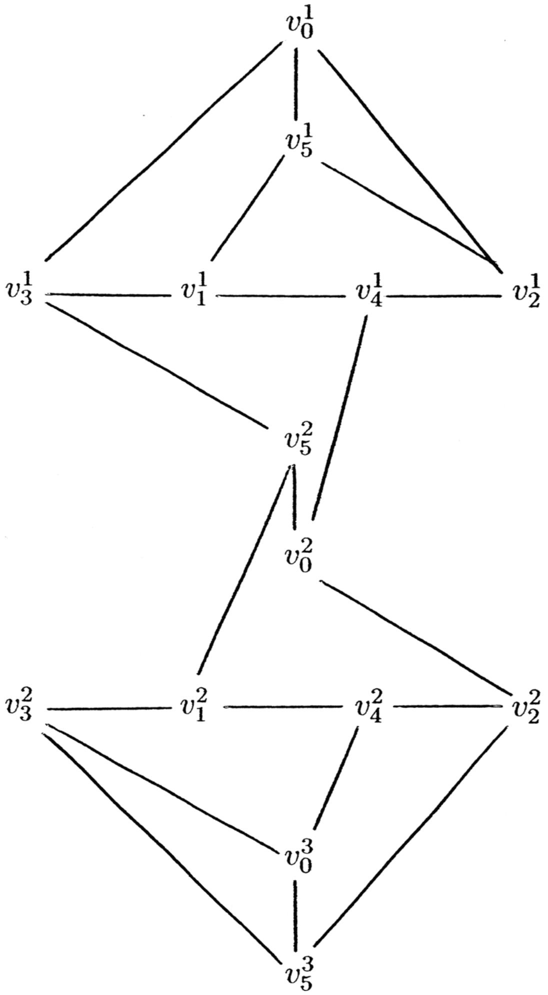


FIG. 5.1. *A $\lambda_{2,1}$-labeling of a $(2,1,3)$-acyclically optimal graph on 14 vertices.*

graphically the ith subset size $r-1$ of $\{1,2,3,\ldots,r\}$. We define the edges of G as follows:

1. For $1 \leq p \leq r$, $\{v_0^p, v_{r+2}^p\} \in E(G)$.
2. For $1 \leq h \leq r-1$, $\{v_s^h, v_t^h\} \in E(G)$ if and only if $|s-t| \geq 2$, $1 \leq s,t \leq r+1$.
3. For $1 \leq p \leq r$, $2 \leq i \leq r+1$ and $1 \leq h \leq r-1$, $\{v_0^p, v_i^h\} \in E(G)$ if and only if S_p contains i and there are exactly $h-1$ sets $S_1, S_2, \ldots, S_{p-1}$ which contain i.
4. For $1 \leq p \leq r$, $1 \leq i \leq r$ and $1 \leq h \leq r-1$, $\{v_{r+2}^p, v_i^h\} \in E(G)$ if and only if T_p contains i and there are exactly $h-1$ sets $T_1, T_2, \ldots, T_{p-1}$ which contain i.

Then the labeling L given by $L(v_z^i) = z$ is a $(2,1,r)$-acyclically optimal labeling of G. In Figure 5.1, we illustrate a $\lambda_{2,1}$-labeling of a $(2,1,3)$-acyclically optimal graph on 14 vertices.

The existence of $(2,1,r)$-optimal graphs on $r+3$ vertices and on r^2+2r-1 vertices leads to the following theorem.

THEOREM 5.4. *For $x, y \in Z^+$, there exists a $(2,1,r)$-optimal graph on $x(r+3)+y(r^2+2r-1)$ vertices.*

COROLLARY 5.5. *If r is even, then for all $n \geq (r+2)(r^2+2r-2)$, there exists a $(2,1,r)$-optimal graph on n vertices. If r is odd, then for all $n \geq \frac{(r-5)(r^2+2r-3)}{4}$, there exists a $(2,1,r)$-optimal graph on $2n$ vertices.*

Proof. If r is even, then $\gcd(r+3, r^2+2r-1) = 1$, implying every integer greater than or equal to $(r+2)(r^2+2r-2)$ can be written as a linear combination of $r+3$ and r^2+2r-1 with nonnegative coefficients. If r is odd, then $\gcd(r+3, r^2+2r-1) = 2$, giving the result by a similar argument. □

Although the $(2,1,r)$-cyclically optimal graph on $r+3$ vertices is unique, such is not the case for $(2,1,r)$-acyclically optimal graphs on r^2+2r-1 vertices. However, each $(2,1,r)$-acyclically optimal graph on r^2+2r-1 vertices, through a sequence of edge transpositions similar to that described in the preceding section, yields a graph isomorphic to $G_a(r)$. Extending this argument gives the following theorem.

THEOREM 5.6. *Every $(2,1,r)$-optimal graph yields, through a sequence of edge transpositions, a graph isomorphic to a sum of copies of $\overline{C_{r+3}}$ and $G_a(r)$.*

6. On prisms. In this section, we apply our results on optimality to a special class of 3-regular graphs known as prisms. For $n \geq 3$, the n-prism, denoted $\mathrm{Pr}(n)$, is the graph consisting precisely of two disjoint n-cycles $v_0, v_1, \ldots, v_{n-1}$ and $w_0, w_1, \ldots, w_{n-1}$ and edges $\{v_i, w_i\}$ for $0 \leq i \leq n-1$. The two cycles shall be called the inner and outer cycles, respectively. We point out that $\mathrm{Pr}(n)$ is isomorphic to $C_n \times P_2$. We also note that it will be convenient to exhibit a labeling of $\mathrm{Pr}(n)$ in the form of a $2 \times n$ array, where the entries in the top row of the array correspond to the labels of the vertices of the outer cycle and the entries in the bottom row correspond to the labels of the vertices of the inner cycle.

In [14], Jha et al. proved the following theorem.

THEOREM 6.1. *Let $n \geq 3$. Then*

$$\lambda_{2,1}(Pr(n)) \begin{cases} = 5 \text{ if } n \equiv 0 \pmod 3, \\ \leq 6 \text{ if } n \not\equiv 0 \pmod 3. \end{cases}$$

We refine this theorem as follows (and are informed that an alternative proof will appear in [15]).

THEOREM 6.2. *Let $n \geq 3$. Then*

$$\lambda_{2,1}(Pr(n)) = \begin{cases} 5 \text{ if } n \equiv 0 \pmod 3, \\ 6 \text{ if } n \not\equiv 0 \pmod 3. \end{cases}$$

Proof. By Theorem 6.1, it suffices to show that $\lambda_{2,1}(\mathrm{Pr}(n)) > 5$ for $n \not\equiv 0 \pmod 3$. Suppose to the contrary that there exists an n, $n \equiv 1, 2 \pmod 3$, such that $\lambda_{2,1}(\mathrm{Pr}(n)) = 5$. Let L be a $\lambda_{2,1}$-labeling of $\mathrm{Pr}(n)$. Since the order of $\mathrm{Pr}(n)$ is $2n$, we observe that $|V(\mathrm{Pr}(n))| \equiv 2, 4 \pmod 6$, which by Theorem 4.3 implies that L is acyclic. Thus, by the discussion following Theorem 5.1, $m(0, i^*) \geq 1$, $2 \leq i \leq 4$, implying $m(0, 3^*) \geq 1$.

With no loss of generality, let $a_0, a_1, a_2, b_0, b_1, b_2$ be vertices in $V(\mathrm{Pr}(n))$ such that $a_1 \in M(0, 3^*)$. Then the neighbors of a_1, namely, a_0, a_2, and b_1, receive the labels 2, 4, and 5 under L (not necessarily respectively). If a_0 or b_1 receives the label 2, then by virtue of the 4-cycle $\langle a_0, a_1, b_1, b_0 \rangle$, $L(b_0) \geq 6$, contradicting the optimality of L. If a_2 receives the label 2, then by virtue of the 4-cycle $\langle a_1, a_2, b_2, b_1 \rangle$, $L(b_2) \geq 6$, another contradiction. □

For $\frac{j}{k} > r$, if $\mathrm{Pr}(n) \in \Gamma(j,k,3)$, then by Theorem 3.1, $\mathrm{Pr}(n)$ is bipartite (implying that n is even) and $2n \equiv 0 \pmod 6$. Hence $n \equiv 0 \pmod 6$, a condition which is easily seen to be sufficient for optimality by labeling the vertices of the inner cycle $0, 2, 4, 0, 2, 4, \ldots, 0, 2, 4$ and the vertices of the outer cycle $3, 5, 1, 3, 5, 1, \ldots, 3, 5, 1$.

If $\mathrm{Pr}(n) \in \Gamma(1,1,3)$, then by Theorem 4.3, $|V(\mathrm{Pr}(n)| \equiv 0 \pmod 4$. Hence, n is even, so $\mathrm{Pr}(n)$ is bipartite. By Theorem 4.5, $|V(\mathrm{Pr}(n)| \equiv 0 \pmod 8$, implying the necessary condition $n \equiv 0 \pmod 4$. However, this condition is also sufficient for the $(1,1,3)$-optimality of $\mathrm{Pr}(n)$, as shown in the definitive calculation of $\lambda_{1,1}(\mathrm{Pr}(n))$, given below.

Theorem 6.3. *Let $n \geq 3$. Then*

$$\lambda_{1,1}(Pr(n)) = \begin{cases} 3 \text{ if } n \equiv 0 \pmod 4, \\ 5 \text{ if } n = 3, 6, \\ 4 \text{ otherwise.} \end{cases}$$

Proof. For $n \equiv 0 \pmod 4$, consider the array A_1, which represents a $\lambda_{1,1}$-labeling of $\mathrm{Pr}(4)$:

$$\begin{matrix} 0\ 1\ 2\ 3 \\ 2\ 3\ 0\ 1. \end{matrix}$$

Then if $n = 4m$, an optimal $(1,1,3)$-labeling of $\mathrm{Pr}(n)$ is demonstrable by catenating m copies of A_1 like so:

$$\begin{matrix} 0\ 1\ 2\ 3\ 0\ 1\ 2\ 3\ \ldots\ 0\ 1\ 2\ 3 \\ 2\ 3\ 0\ 1\ 2\ 3\ 0\ 1\ \ldots\ 2\ 3\ 0\ 1. \end{matrix}$$

For $n = 3, 6$, consider $n = 3$. Then $\mathrm{Pr}(3)$ has diameter two, which implies (by the distance conditions) that no two vertices may be assigned the same label. It is then an easy matter to show the existence of an $L(1,1)$-labeling of $\mathrm{Pr}(3)$ with span equal to the lower bound 5. If $n = 6$, then the converse of part a implies that $\lambda_{1,1}(\mathrm{Pr}(6)) \geq 4$. But if $\lambda_{1,1}(\mathrm{Pr}(6)) = 4$, the pigeon-hole principle implies the existence of a label with multiplicity 3. The distance constraints, however, imply that no label may have multiplicity 3. Thus, $\lambda_{1,1}(\mathrm{Pr}(6)) = 5$, as demonstrated by the following labeling:

$$\begin{matrix} 0\ 1\ 2\ 3\ 4\ 5 \\ 2\ 3\ 4\ 5\ 0\ 1. \end{matrix}$$

In the final case, we note that n is not a multiple of 4, implying that $\lambda_{1,1}(\mathrm{Pr}(n)) \geq 4$. It thus suffices to show the existence of an $L(1,1)$-labeling with span 4. To that end, consider the array A_2, which represents a $\lambda_{1,1}$-labeling of $\mathrm{Pr}(5)$:

$$\begin{matrix} 0\ 1\ 2\ 3\ 4 \\ 2\ 3\ 4\ 0\ 1. \end{matrix}$$

Then, since any integer n, $n > 11$ and n not divisible by 4, can be written $4\alpha + 5\beta$ for some $\alpha \geq 0$ and $\beta \geq 1$, we can demonstrate an $L(1,1)$-labeling with span 4 by the catenating α copies of A_1 and β copies of A_2.

In the remaining cases $n = 7$ and $n = 11$, we demonstrate $L(1,1)$-labelings with span 4:

$$\begin{matrix} 0\ 4\ 1\ 0\ 3\ 1\ 2 \\ 1\ 2\ 3\ 4\ 2\ 0\ 3 \end{matrix}$$

and

$$\begin{array}{l} 0\ 2\ 3\ 1\ 0\ 2\ 4\ 1\ 0\ 3\ 4 \\ 3\ 1\ 0\ 4\ 3\ 1\ 0\ 2\ 4\ 1\ 2. \qquad \square \end{array}$$

Let the Cartesian product of the infinite path P_∞ and P_2 be denoted by $\mathrm{Pr}(\infty)$. By an approach similar to the one used in the proof of Theorem 2.3, we may establish that $\lambda_{j,k}(\mathrm{Pr}(\infty)) \leq \lambda_{j,k}(\mathrm{Pr}(n))$ for all $n \geq 3$; furthermore, it can be shown that $\lambda_{1,1}(\mathrm{Pr}(\infty)) = 3$ and $\lambda_{2,1}(\mathrm{Pr}(\infty)) = 5$ and that $\mathrm{Pr}(\infty)$ is both $(1,1,3)$- and $(2,1,3)$-cyclically optimal. (By Theorem 4.1, $\mathrm{Pr}(\infty)$ is $(3,1,3)$-optimal.) Finally, analysis analogous to that employed in the proof of Theorem 6.2 reveals that all optimal $(2,1,3)$-labelings of $\mathrm{Pr}(\infty)$ are cyclic with even labels appearing along one copy of P_∞ and odd labels along the other.

REFERENCES

[1] G. J. Chang, W.-T. Ke, D. Kuo, D. Liu, and R. Yeh, *On $L(d,1)$-labelings of graphs*, Discrete Math., 220 (2000), pp. 57–66.

[2] G. J. Chang and D. Kuo, *The $L(2,1)$-labeling problem on graphs*, SIAM J. Discrete Math., 9 (1996), pp. 309–316.

[3] G. Chartrand, D.J. Erwin, F. Harary, and P. Zhang, *Radio labelings of graphs*, Bull. Inst. Combin. Appl., 33 (2001), pp. 77–85.

[4] J. P. Georges and D. W. Mauro, *Labeling trees with a condition at distance two*, Discrete Math., 269 (2003), pp. 127–148.

[5] J. P. Georges and D. W. Mauro, *Some results on the $\lambda_{j,k}$-numbers of products of complete graphs*, Congr. Numer., 140 (1999), pp. 141–160.

[6] J. P. Georges and D. W. Mauro, *On the size of graphs labeled with a condition at distance two*, J. Graph Theory, 22 (1996), pp. 47–57.

[7] J. P. Georges and D. W. Mauro, *Generalized vertex labelings with a condition at distance two*, Congr. Numer., 109 (1995), pp. 141–159.

[8] J. P. Georges, D. W. Mauro, and M. I. Stein, *Labeling products of complete graphs with a condition at distance two*, SIAM J. Discrete Math., 14 (2000), pp. 28–35.

[9] J. P. Georges, D. W. Mauro, and M. A. Whittlesey, *Relating path coverings to vertex labelings with a condition at distance two*, Discrete Math., 135 (1994), pp. 103–111.

[10] J. P. Georges and D. W. Mauro, *On the criticality of graphs labeled with a condition at distance two*, Congr. Numer., 101 (1994), pp. 33–49.

[11] J. R. Griggs and R. K. Yeh, *Labelling graphs with a condition at distance* 2, SIAM J. Discrete Math., 5 (1992), pp. 586–595.

[12] W. K Hale, *Frequency assignment: Theory and application*, Proc. IEEE, 68 (1980), pp. 1497–1514.

[13] P. Jha, *Optimal $L(2,1)$-Labeling of Kronecker Products of Certain Cycles*, preprint, Universiti Multimedia Telekom, Malaysia.

[14] P. Jha, A. Narayanan, P. Sood, K. Sundaram, and V. Sunder, *$L(2,1)$-labeling of the Cartesian product of a cycle and a path*, Ars Combin., 55 (2000), pp. 81–89.

[15] D. Kuo and J. H. Yan, *On $L(2,1)$-labelings of Cartesian products of paths and cycles*, Discrete Math., to appear.

[16] D. Liu and R. Yeh, *On distance two labelings of graphs*, Ars Combin., 47 (1997), pp. 13–22.

[17] D. Sakai, *Labeling chordal graphs: Distance two condition*, SIAM J. Discrete Math., 7 (1994), pp. 133–140.

[18] J. Van Den Heuvel, R. A. Leese, and M. A. Shepherd, *Graph labeling and radio channel assignment*, J. Graph Theory, 29 (1998), pp. 263–283.

[19] M. A. Whittlesey, J. P. Georges, and D. W. Mauro, *On the λ-number of Q_n and related graphs*, SIAM J. Discrete Math., 8 (1995), pp. 499–506.

[20] P. Zhang, *Radio labelings of cycles*, Ars Combin., 65 (2002), pp. 21–32.

SIAM J. DISCRETE MATH.
Vol. 17, No. 2, pp. 332–340

ADDITIVE TREE SPANNERS*

DIETER KRATSCH[†], HOÀNG-OANH LE[‡], HAIKO MÜLLER[§], ERICH PRISNER[¶], AND DOROTHEA WAGNER[‖]

Abstract. A spanning tree of a graph is a *k-additive tree spanner* whenever the distance of every two vertices in the graph differs from that in the tree by at most k. In this paper we show that certain classes of graphs, such as distance-hereditary graphs, interval graphs, asteroidal triple-free graphs, allow some constant k such that every member in the class has some k-additive tree spanner. On the other hand, there are chordal graphs without a k-additive tree spanner for arbitrarily large k.

Key words. distance, graph spanner, spanning tree, algorithm, asteroidal triple-free graph, distance-hereditary graph, chordal graph

AMS subject classifications. 05C12, 05C85, 68R10, 68Q20, 90B80

DOI. 10.1137/S0895480195295471

1. Introduction.

Spanning trees are often used in applications where we want to save edges but maintain connectivity. Since in most of these applications distance matters, not all spanning trees have the same quality. It is desirable that vertices of small distance in the original graph should also have relatively small distance in the spanning tree. If we require $d_T(x,y)/d_G(x,y) \leqslant k$ for all pairs of vertices x, y, then we arrive at the well-known concept of k-multiplicative tree spanners; compare [3]. The stronger property $d_T(x,y) - d_G(x,y) \leqslant k$ for every pair of vertices x, y, defines *k-additive* spanners; see [14]. In this paper we shall deal with spanning trees that are k-additive spanners. We call them *k-additive tree spanners*.

Spanners have received a lot of attention in the last few years. They have been introduced by Peleg and Ullman [18] for the purpose of synchronizing asynchronous networks. Spanners have many applications, for instance, in communication networks [18], broadcasting, routing, or robotics. We refer to the papers [17, 20, 21] for more information.

As a kind of extreme case of spanners, tree spanners occur mostly in applications where the cost is the main concern, or where the tree structure is exploited. The question to decide whether a given graph has some k-multiplicative tree spanner is NP-complete for each fixed $k \geqslant 4$, and it can be decided in polynomial-time for $k = 2$ [3, 4]. Finding the smallest k for which a graph has some k-multiplicative tree spanner is also NP-complete even for planar graphs [8]. See also [15, 2, 12] for further information on (multiplicative) tree spanners in special families of graphs.

We will show that certain well-known graph classes Γ allow some constant k_Γ such that every graph in Γ has some k_Γ*-additive* tree spanner. We present two simple

[*] Received by the editors December 1, 1995; accepted for publication (in revised form) April 17, 2003; published electronically December 19, 2003.
http://www.siam.org/journals/sidma/17-2/29547.html

[†] LITA, Université dé Metz, Metz, France (kratsch@sciences.univ-metz.fr).

[‡] Fachbereich Informatik, Universität Rostock, Rostock, Germany (hoang-oanh.le@informatik.uni-rostock.de).

[§] School of Computing, University of Leeds, Leeds, UK (hm@comp.leeds.ac.uk).

[¶] Department of Mathematics, Franklin College, Lugano, Switzerland (eprisner@fc.edu). The work of this author was supported by the Deutsche Forschungsgemeinschaft under grant Pr 324/6-1.

[‖] Fakultät für Mathematik und Informatik, Universität Konstanz, Konstanz, Germany (dorothea.wagner@uni-konstanz.de).

approaches for finding such trees. The first one is to consider certain breadth-first search trees. For interval graphs and distance-hereditary graphs, we get 2-additive tree spanners in linear time. The second approach requires that we have some dominating shortest path. We connect the vertices outside the path to the path in a consistent way to obtain a tree spanner, which, as can be shown, is 4-additive.

Some other classes Γ of graphs do not have such a constant k_Γ. This holds for instance if all cycles belong to Γ, since cycles of length $k+3$ do not have a k-additive tree spanner. But even the class of chordal graphs does not have such a constant, as we shall see.

The fundamental notion in the paper is the *distance* $d_G(x,y)$ between two vertices x and y in the connected graph G. It is defined to be the length (number of edges) of a shortest x-y path. All graphs occurring in this paper are supposed to be connected.

Parts of the results of the paper appeared in [11] and [19].

2. Spanning trees growing from isometric subtrees. For a set $W \subseteq V$ of vertices in a graph G and any integer $i \geqslant 0$ let $N^i(W)$ denote the ith neighborhood of W, i.e., the set of all vertices y with $d_G(y,W) = i$, where $d_G(y,W) = \min_{w\in W} d_G(y,w)$.

DEFINITION 2.1. *A subtree T of a graph $G = (V,E)$ is an* isometric subtree *if $d_G(u,v) = d_T(u,v)$ for all $u,v \in V(T)$.*

Isometric subtrees are necessarily induced subgraphs.

The most straightforward method of constructing a good additive tree spanner in a connected graph G seems to be the following.

```
Basic construction:
begin
        choose a set $W \subseteq V$, $W \neq \varnothing$, such that $G[W]$ is an isometric subtree of $G$;
        $T \leftarrow G[W]$; $i \leftarrow 1$;
        while $N^i(W) \neq \varnothing$ do begin
                for $y \in N^i(W)$ do begin
                        choose a vertex $z = f(y) \in N(y) \cap N^{i-1}(W)$;
                        $V(T) \leftarrow V(T) \cup \{y\}$;
                        $E(T) \leftarrow E(T) \cup \{yz\}$;
                end;
                $i \leftarrow i+1$;
        end;
end.
```

Since G is connected this construction leads to a spanning tree $T = (V(T), E(T))$ of G. Then every $y \in N^i(W), i \geqslant 1$, has exactly one neighbor $f(y)$ in $N^{i-1}(W)$ on the tree constructed. So, what we do is construct a spanning tree T where all distances from W toward the other vertices are identical in G and T.

LEMMA 2.2. *If $G[W]$ is an isometric subtree of G such that $N^{k+1}(W) = \varnothing$, then a spanning tree T of G constructed by our basic construction is $4k$-additive.*

Proof. We consider a shortest path $(v_0, v_1, \dots, v_t)$ in G. Let w_0 and w_t be those vertices in W such that $d_T(v_i,w_i) = d_G(v_i,W)$ for $i = 0,t$. Then $d_T(v_i,w_i) \leqslant k$ since $N^{k+1}(W) = \varnothing$ and $d_T(w_0,w_t) \leqslant d_G(w_0,v_0) + d_G(v_0,v_t) + d_G(v_t,w_t) \leqslant 2k+t$ since $G[W]$ is an isometric subtree of G. Consequently, $d_T(v_0,v_t) \leqslant 4k+t = 4k + d_G(v_0,v_t)$. □

The algorithm still contains two ambiguities. How do we choose our starting subtree $G[W]$, and which function $f : V \setminus W \to V$ (selecting the neighbors closer to W) do we use? As isometric subtree $G[W]$ we will use either a single vertex (in sections 3 and 4) or a dominating shortest path of G (in section 5). Later we will describe how to define f (for arbitrary graph G) using some "rules."

Let us now consider the special case $W = \{x_0\}$. In order to find out whether such a resulting tree is k-additive, we do not have to compute all T-distances, but only T-distances between G-adjacent vertices.

LEMMA 2.3. *A spanning tree T of G constructed by our basic construction starting from a singleton x_0 is k-additive if and only if $d_T(y,z) \leqslant k+1$ for every edge yz of G.*

Proof. Necessity of this condition is obvious.

For sufficiency, assume that $d_T(y,z) \leqslant k+1$ for every edge yz of G. By induction over t we prove $d_T(v_0, v_t) \leqslant k+t$ for paths $v_0, v_1, \ldots, v_t$ of length t in G. The case $t = 1$ is just the assumption; now assume that $d_G(u,v) = t > 1$, and assume that $d_T(y,z) \leqslant d_G(y,z) + k$ whenever $d_G(y,z) < t$ (the induction hypothesis). Choose any vertex w on a shortest u-v path. Let $T(u,v,w)$ be the smallest subtree of T containing the vertices u, v, and w. Let p be the vertex of $T(u,v,w)$ separating u, v, and w from each other. Note that p would be one of the vertices u, v, and w if $T(u,v,w)$ would be a path. Let q be the vertex of $T(u,v,w)$ having smallest distance to the root x_0. We distinguish two cases.

Case 1 (q lies on the u-v path in T). By definition of p, p has to lie on this path too. Since the situation is symmetric in u and v, we may assume that p lies on the u-q path. We get

$$\begin{aligned} d_T(u,v) &= d_T(u,p) + d_T(p,q) + d_T(q,v) \\ &= (d_T(u,p) - d_T(w,p)) + (d_T(w,p) + d_T(p,q) + d_T(q,v)) \\ &= (d_T(u,p) - d_T(w,p)) + d_T(w,v). \end{aligned}$$

By the construction of T,

$$d_T(u,p) - d_T(w,p) = d_T(u,x_0) - d_T(w,x_0) = d_G(u,x_0) - d_G(w,x_0) \leqslant d_G(u,w)\,.$$

For the second term we apply the induction hypothesis for the vertices w and v of G-distance smaller than $d_G(u,v)$ to get

$$d_T(u,v) \leqslant d_G(u,w) + d_T(w,v) \leqslant d_G(u,w) + d_G(w,v) + k = d_G(u,v) + k\,.$$

Case 2 (q does not lie on the T-path between u and v). Applying the induction hypothesis to the pairs u, w and w, v, both of G-distance smaller than $d_G(u,v)$, we get

$$\begin{aligned} d_T(u,p) + d_T(p,q) + d_T(q,w) = d_T(u,w) \leqslant d_G(u,w) + k, \\ d_T(v,p) + d_T(p,q) + d_T(q,w) = d_T(v,w) \leqslant d_G(v,w) + k. \end{aligned}$$

As in Case 1, we get

$$\begin{aligned} d_T(u,p) + d_T(p,q) - d_T(q,w) \leqslant d_G(u,w), \\ d_T(v,p) + d_T(p,q) - d_T(q,w) \leqslant d_G(v,w). \end{aligned}$$

Adding the previous four inequalities yields

$$2d_T(u,p) + 2d_T(v,p) + 4d_T(p,q) \leqslant 2(d_G(u,w) + d_G(v,w) + k)$$

and therefore

$$d_T(u,v) = d_T(u,p) + d_T(v,p) \leqslant d_G(u,w) + d_G(v,w) + k = d_G(u,v) + k\,. \qquad \square$$

Although starting with singletons has this strong property and works for several graph classes, as we shall see, it also has one disadvantage. It does not appear to construct the *optimum* spanning tree—that is, one which is k-additive for the smallest possible k.

As an example, the graph of Figure 1 has some 1-additive tree spanner, which, however, will not be found by our approach under *any* rule and *any* start vertex x_0.

On the other hand, if we start our construction with maximum isometric subtrees, then the assertion of Lemma 2.3 does not hold. As an example, on the graph G depicted in Figure 2 we obtain a spanning tree T with $d_T(y,z) \leqslant 4$ for every edge yz of G if we start with a maximum isometric subtree (all such subtrees are shortest paths on five vertices). However, T is not 3-additive.

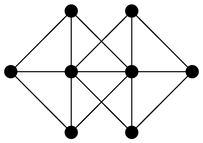


FIG. 1. $|W| = 1$ *is not optimal.*

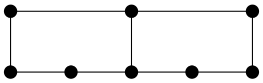


FIG. 2. $|W| = \max$ *is not optimal.*

For the algorithms actually constructing these tree spanners, we assume that the vertices of G are linearly ordered, and that the graph is given by means of its ordered adjacency lists, i.e., for every vertex x there is some list NEIGH(x) containing its neighbors in increasing order.

The levels $N^i(W)$ can be computed in linear time by breadth-first search. Moreover, it is also possible to compute the induced subgraphs $G[N^i(W)]$ in linear time—we simply remove from the list NEIGH(x) all neighbors of x contained in a level different from the one containing x itself. Doing so for all vertices x, we obtain the graph $\bigcup_i G[N^i(W)]$.

How quickly T can be constructed surely depends on the rules, which themselves depend on the graph classes considered.

3. Distance-hereditary graphs.

DEFINITION 3.1. *A connected graph $G = (V,E)$ is* distance-hereditary *if every induced x-y path, $x,y \in V$, has length $d_G(x,y)$.*

Several characterizations of distance-hereditary graphs, which are sometimes called "completely separable graphs," are given in [1] and [9]. For a fixed vertex x_0 in a distance-hereditary graph, two vertices y and z are *tied* if there is some vertex w and some shortest w-x_0 path containing y and some shortest w-x_0 path containing z.

LEMMA 3.2 (see [9]). *For every distance-hereditary graph $G = (V,E)$, for every $x_0 \in V$, and for every nonnegative integer i, every two adjacent or tied vertices in $N^{i+1}(x_0)$ have the same neighbors in $N^i(x_0)$.*

Let x_0 be any fixed vertex in a connected distance-hereditary graph G. Given an arbitrary linear order of the vertices of G, we define the following rule.

Rule 1. Connect every $y \in N^{i+1}(x_0)$ with the smallest (with respect to that order) vertex $f(y) \in N(y) \cap N^i(x_0)$.

Applying Rule 1, we construct a breadth-first search tree of G. Thereby we start at x_0 and scan the unvisited neighbors of the vertex under consideration in the specified order.

Theorem 3.3. *Every connected distance-hereditary graph G has some 2-additive tree spanner which can be found by our basic construction and Rule 1 in linear time.*

Proof. Let T be the tree constructed by the rule above. By Lemma 2.3 we may restrict our attention to a single edge yz of G.

(1) If $d_G(y, x_0) = d_G(z, x_0) = i+1$, then $N(y) \cap N^i(x_0) = N(z) \cap N^i(x_0)$ by Lemma 3.2; thus $f(y) = f(z)$, and $d_T(y, z) \leqslant 2$.

(2) If w.l.o.g. $i + 1 = d_G(y, x_0) = d_G(z, x_0) + 1$, we may assume $f(y) \neq z$—otherwise we are already done.

Then there are shortest y-x_0 paths in G, one going over $f(y)$, and one over z. Then $f(y)$ and z are tied, therefore $N(f(y)) \cap N^{i-1}(x_0) = N(z) \cap N^{i-1}(x_0)$ by Lemma 3.2. Therefore $f(f(y)) = f(z)$, and $d_T(y, z) = 3$, as desired.

Finding the tree T is very easy: We simply compute the levels and then check for every vertex y the adjacency list in increasing order until we find some vertex one level beyond the level of y—the resulting vertex is $f(y)$. □

In Figure 3 at the end of section 4, we will give a distance-hereditary graph without a 1-additive tree spanner showing that the bound given in Theorem 3.3 is optimal.

For the subclass of block graphs, the above construction with *any* rule yields a 1-additive tree spanner [19].

4. Interval graphs.

Definition 4.1. *A graph is an* interval graph *if one can associate with each vertex an interval on the real line such that two vertices are adjacent if and only if the corresponding intervals have a nonempty intersection.*

In [13] the interval graphs are characterized as those chordal graphs (see Definition 6.1 on page 338) without asteroidal triples (see Definition 5.1 on page 337).

We rely on the following property of chordal graphs, which is not hard to prove.

Lemma 4.2. *Let x_0 be a vertex in the chordal graph G.*

(a) *For every $x \in N^{i+1}(x_0)$, $N(x) \cap N^i(x_0)$ induces a complete graph.*

(b) *For every edge $yz \in N^{i+1}(x_0)$, the two sets $N(y) \cap N^i(x_0)$ and $N(z) \cap N^i(x_0)$ must be comparable by set inclusion.*

Proof. (a) Assume $x \in N^{i+1}(x_0)$ had two nonadjacent neighbors y, z in $N^i(x_0)$. Then we choose some chordless y-z path that, except y and z, uses only vertices inside the levels $N^0(x_0)$ up to $N^{i-1}(x_0)$. Together with the edges zx and xy it forms an induced cycle of length 4 or more, a contradiction.

(b) Again, assume to the contrary that there are vertices $y', z' \in N^i(x_0)$ such that y' is adjacent to y but not to z, and z' is adjacent to z but not to y. If y' and z' are adjacent, then we have some induced 4-cycle in G. Otherwise, again we find some induced y'-z' path where all internal vertices have distance less than i to x_0. Together with the edges $z'z, zy, yy'$ this path yields an induced cycle of length at least 5 in G, a contradiction again. □

Lemma 4.3. *For every interval graph and every connected component Q of $G[N^{i+1}(x_0)]$, there is some vertex $f(Q) \in N^i(x_0)$ adjacent to all vertices of Q.*

Proof. We assume that there are two vertices y and z in some common connected component Q of $G[N^{i+1}(x_0)]$, whose sets of neighbors inside $N^i(x_0)$ are not comparable. By Lemma 4.2(b), y and z are not adjacent. We find y' and z' in $N^i(x_0)$ such that y' is adjacent to y but not to z, and z' is adjacent to z but not to y.

Then the three vertices x_0, y, and z form an asteroidal triple (see Definition 5.1): The y-z path inside $N^{i+1}(x_0)$ avoids x_0 and its neighbors, every shortest y-x_0 path going through y' avoids z and its neighbors, and every shortest z-x_0 path going over

z' avoids y and its neighbors. Consequently the nonempty sets $N_G(y) \cap N^i(x_0), y \in V(Q)$, form a chain; thus some element is contained in all these sets. But an interval graph cannot contain asteroidal triples [13]. □

Let all the vertices $f(Q)$ for the components Q of the levels $G[N^{i+1}(x_0)]$ be chosen in advance; for instance, we could choose the smallest element (in the given ordering of the vertices) in $\bigcap_{y \in Q} N^i(y) \cap N^i(x_0)$. The following rule fits into our general scheme.

Rule 2. Connect $y \in N^{i+1}(x_0)$ with $f(Q)$, where Q denotes the component of $G[N^{i+1}(x_0)]$ containing y.

Now we are able to present our approach for interval graphs. Note that 2-additive tree spanners are also found by a different algorithm in [15].

THEOREM 4.4. *Every connected interval graph has some* 2*-additive tree spanner, that can be found by our basic construction and Rule* 2 *in linear time.*

Proof. By Lemma 2.3 it suffices to show that $d_T(y, z) \leqslant 3$ for every edge yz of G for the tree T constructed in this way. The case where both y and z have the same distance $d_G(y, x_0) = d_G(z, x_0) = i + 1$ toward x_0 is easy: Then both lie in the same component Q of $G[N^{i+1}(x_0)]$, whence they are connected over $f(Q)$ in T. So assume that $i+1 = d_G(y, x_0) = d_G(z, x_0)+1$, and let Q denote the component of $G[N^{i+1}(x_0)]$ containing y. Since we are done if $f(Q) = z$, assume $f(Q) \neq z$. Then both $f(Q)$ and z are neighbors of y in $N^i(x_0)$—by Lemma 4.2(a) they must be adjacent. By the construction z and $f(Q)$ have distance 2 in T; thus $d_T(y, z) \leqslant 3$.

To find T, we first run a breadth-first search on G and obtain levels $N^i(x_0)$. That is, we mark every vertex x by the label $d_G(x, x_0)$. Then we compute the components of the graphs $G[N^i(x_0)]$ for all levels i in linear time as mentioned in section 2. Then the algorithm identifies $f(Q)$ by counting for each vertex of level $i - 1$ and each component Q the number of common edges. Since the collection of all components Q of graphs $G[N^i(x_0)]$ for all i forms a partition of the vertex set of G, every adjacency list is traversed exactly once, showing that the overall running time of our algorithm is linear. □

Both Theorems 4.4 and 3.3 are best possible. The graph in Figure 3 has no 1-additive tree spanner, but it is both a distance-hereditary and an interval graph.

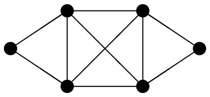


FIG. 3. *A graph without* 1*-additive tree spanner.*

5. Asteroidal triple-free graphs.

DEFINITION 5.1. *An independent set of three vertices is called an* asteroidal triple *if between each pair in the triple there exists a path that avoids the neighborhood of the third. A graph is asteroidal triple-free (AT-free) if it contains no asteroidal triple.*

In this section we use a dominating shortest path (DSP) as an isometric subtree. More precisely, a shortest path $(x_0, x_1, \ldots, x_\ell)$ in $G = (V, E)$ is *dominating* if every vertex in $V \setminus W$, $W = \{x_0, x_1, \ldots, x_\ell\}$, is adjacent to at least one vertex in W. By Lemma 2.2 we know that every graph with DSP has a 4-additive tree spanner. This applies to all AT-free graphs as a consequence of the dominating pair theorem given in [6]. It is worth mentioning that a DSP in an AT-free graph G can be found in linear time by 2 × LexBFS [5]: First start a lexicographic breadth-first search (LexBFS) from an arbitrary vertex x of G. Let x_0 be the vertex numbered last by this search. Start a second LexBFS in G from x_0, and let x_ℓ ($\ell = d_G(x_0, x_\ell)$) be the last vertex in the second LexBFS. In [5] it is shown that *every* shortest path $(x_0, x_1, \dots, x_\ell)$ is a DSP of G.

Next we demonstrate how to use such a DSP in an AT-free graph to show that every connected AT-free graph admits a 3-additive tree spanner. We need the following result from [10].

LEMMA 5.2. *The* DSP $(x_0, x_1, \dots , x_\ell)$, *constructed as in the proof of Theorem* 7 *in* [10] *for a given connected AT-free graph G=(V,E) in linear time, has the following property: For every* $i = 1, 2, \dots, \ell$, *every vertex* $z \in N^i(x_0)$ *is adjacent to* x_i *or* x_{i-1}.

Rule 3. For $i = 1, 2, \dots, \ell$ connect a vertex $v \in N^i(x_0)$ to $f(v) = x_i$ if v is adjacent to x_i; otherwise connect v to $f(v) = x_{i-1}$.

THEOREM 5.3. *Every connected AT-free graph has a* 3*-additive tree spanner that can be found in linear time.*

Proof. Let $G = (V, E)$ be a connected AT-free graph, and let $(x_0, x_1, \dots, x_\ell)$ be a DSP of G constructed by $2 \times$LexBFS. To construct a 3-additive tree spanner of G, we use our basic construction with $G[W]$ as an isometric subtree, $W = \{x_0, x_1, \dots, x_\ell\}$, and apply Rule 3. This defines a spanning tree T of G, since by Lemma 5.2 for all $v \in V \setminus W$ the vertex $f(v) \in W$ is adjacent to v.

We consider a shortest path $(v_0, v_1, \dots, v_t)$ in G with $t \geqslant 1$ (Lemma 2.3 does not apply), and w.l.o.g. $v_0 \in N^i(x_0)$ and $v_t \in N^j(x_0)$, $0 \leqslant i \leqslant j \leqslant \ell$. Then $j - i \leqslant t$ since $d_G(x_0, v_t) \leqslant d_G(x_0, v_0) + d_G(v_0, v_t)$. This implies $d_T(f(v_0), f(v_t)) \leqslant t + 1$ by Rule 3. Finally, since $(x_0, x_1, \dots, x_\ell)$ is a DSP of G, we have $d_T(v_0, v_t) \leqslant t+3 = d_G(v_0, v_t)+3$.

Observe that the 3-spanner T constructed in the proof can be found in linear time, since we can find a DSP in AT-free graphs by 2 × LexBFS [5]. □

Moreover, Theorem 5.3 gives the best possible bound since the 5-cycle is an AT-free graph which has no 2-additive tree spanner.

A graph $G = (V, E)$ is a *cocomparability graph* if there is some poset $(V, <)$ such that distinct vertices are adjacent in G if and only if they are not comparable. It is well known that the cocomparability graphs form a class between the interval graphs and the AT-free graphs [13]. Hence we know by Theorem 5.3 that every cocomparability graph has a 3-additive tree spanner. However, the chordless 4-cycle which has no 1-additive tree spanner gives the best-known lower bound for this class.

Conjecture. Every cocomparability graph admits a 2-additive tree spanner.

In the earlier mentioned algorithm for interval graphs by Madanlal et al. [15], the above approach is used implicitly.

6. Chordal graphs.

DEFINITION 6.1. *A graph is called* chordal *if it does not contain a chordless cycle of length greater than* 3.

Since block graphs, as well as interval graphs, are chordal, considering this class might seem promising. Also, n-vertex chordal graphs have 2-multiplicative spanners with $O(n^{1.5})$ edges, 3-multiplicative spanners with $O(n \log n)$ edges, and 5-multiplicative spanners with $2n - 2$ edges [17]. However, the following example, independently found by McKee [16] (see also [11]), shows that for every fixed integer k there are chordal graphs without k-additive *tree* spanners.

Let the graph G_1 be the triangle K_3, and let G_2 be the graph obtained from G_1 by adding three vertices, each one adjacent to two distinct vertices of G_2. Let, for every integer $s \geqslant 2$, the graph G_s be obtained from G_{s-1} and G_{s-2} by adding for every edge in $E(G_{s-1}) \setminus E(G_{s-2})$ one new vertex, adjacent to the two vertices of the edge. These graphs are planar (even outerplanar) and chordal (even 2-trees and path

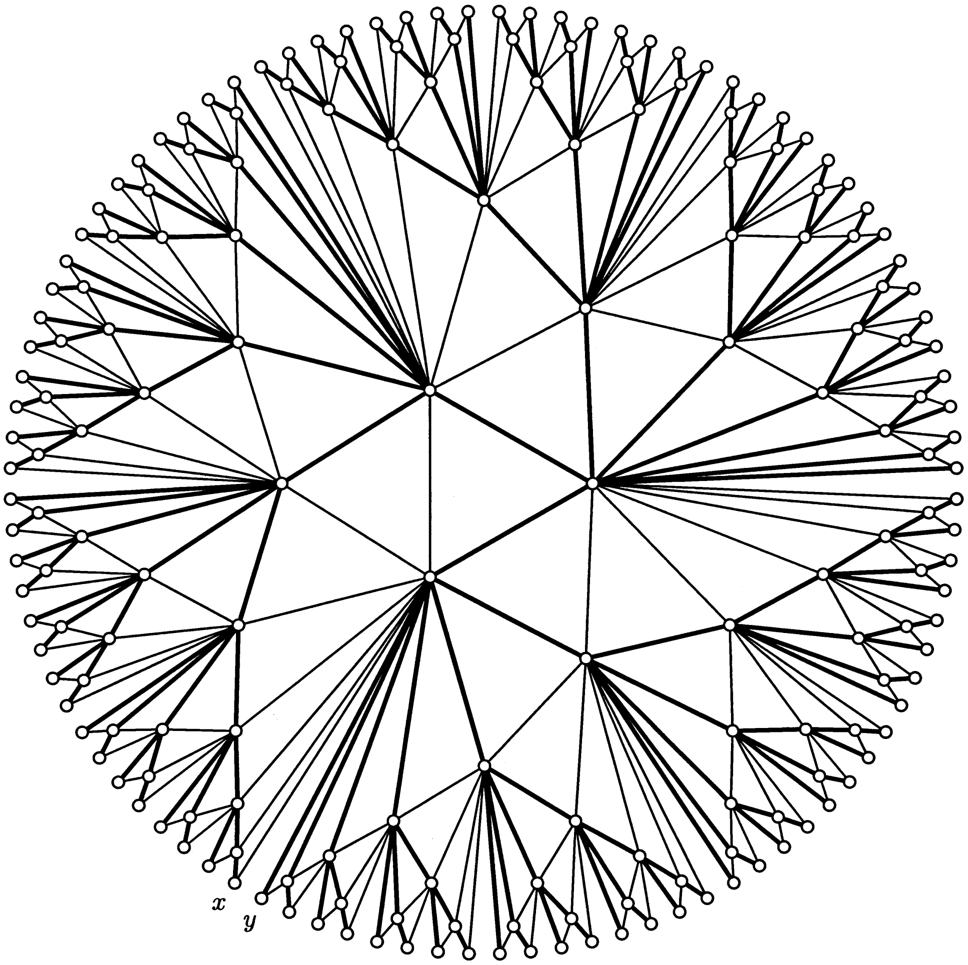


FIG. 4. *A chordal graph with spanning tree indicated by bold lines.*

graphs). The graph G_7, together with some 7-additive tree spanner, is given in Figure 4. Look at vertices x and y to see that this particular tree is not 6-additive.

PROPOSITION 6.2. *No $(k-1)$-additive tree spanner and no k-multiplicative tree spanner is possible in G_k.*

Proof. The *eccentricity* $\mathrm{ecc}_G(x)$ of x is the largest integer i for which $N^i(x)$ is nonempty (i.e., $\mathrm{ecc}_G(x) = \max_{y \in V(G)} d_G(x,y)$).

Look at the canonical embedding of G_k in the plane. The first observation is that the outer face F_0 of G_k has, as vertex in the dual G_k^*, eccentricity equal to k. In fact, all faces have the same eccentricity k in the dual graph for this example.

Let T be a spanning tree of G_k. The dual tree T^* contains all edges of G_k^* which cross edges of G_k that do *not* belong to T.

Let B be a largest connected component of the forest $T^* - F_0$, and let F_1 be the neighbor of F_0 in B. Note that B contains at least $\mathrm{ecc}_{T^*}(F_0) \geqslant \mathrm{ecc}_{G_k{}^*}(F_0) = k$ vertices.

The edge F_0F_1 in T^* crosses an edge xy on the outer cycle in G_k. Since F_0F_1

is an edge in T^*, x and y are not adjacent in T. Moreover, $d_T(x,y)+1$ equals the number of edges in G_k^* that start in B and end outside B. Since G_k is outerplanar, the only edges in G_k^* between vertices of B are the edges of T^*. Therefore

$$d_T(x,y)+1 = \sum_{F\in V(B)} (d_{G_k^*} - d_{T^*[B]}(F)) = \sum_{F\in V(B)} (3 - d_{T^*[B]}(F))\,,$$

since all vertices except F_0 have degree 3 in G_k^*. But by the well-known degree sum formula, and since $T^*[B]$ is a tree, this equals

$$3|V(B)| - 2|E(B)| = |V(B)| + 2\,,$$

a number which is greater or equal to $k+2$. Therefore $d_T(x,y) \geqslant k+1$, and T cannot be $(k-1)$-additive or k-multiplicative. □

The class of *strongly chordal graphs* is between the chordal graphs and the interval graphs; see [7] for a definition and characterizations. Strongly chordal graphs allow 3-additive spanning trees, as has been shown recently by a completely different approach [2].

REFERENCES

[1] H.-J. Bandelt and H. M. Mulder, *Distance-hereditary graphs*, J. Combin. Theory Ser. B, 41 (1986), pp. 182–208.

[2] A. Brandstädt, V. Chepoi, and F. Dragan, *Distance approximating trees for chordal and dually chordal graphs*, J. Algorithms, 30 (1999), pp. 166–184.

[3] L. Cai and D. G. Corneil, *Tree spanners: An overview*, Congr. Numer., 88 (1992), pp. 65–76.

[4] L. Cai and D. G. Corneil, *Tree spanners*, SIAM J. Discrete Math., 8 (1995), pp. 359–387.

[5] 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.

[6] D. G. Corneil, S. Olariu, L. Stewart, *Asteroidal triple-free graphs*, SIAM J. Discrete Math., 10 (1997), pp. 399–430.

[7] M. Farber, *Characterizations of strongly chordal graphs*, Discrete Math., 43 (1983), pp. 173–189.

[8] S. P. Fekete and J. Kremer, *Tree spanners in planar graphs*, in Graph-Theoretic Concepts in Computer Science, Lecture Notes in Comput. Sci. 1517, Springer, Berlin, 1998, pp. 298–309.

[9] P. L. Hammer and F. Maffray, *Completely separable graphs*, Discrete Appl. Math., 27 (1990), pp. 85–99.

[10] T. Kloks, D. Kratsch, and H. Müller, *Approximating the bandwidth for asteroidal triple-free graphs*, J. Algorithms, 32 (1999), pp. 41–57.

[11] Hoàng-Oanh Le, *Effiziente Algorithmen für Baumspanner in chordalen Graphen*, Diploma thesis, TU Berlin, Berlin, 1994.

[12] Hoàng-Oanh Le and Van Bang Le, *Optimal tree 3-spanners in directed path graphs*, Networks, 34 (1999), pp. 81–87.

[13] C. Lekkerkerker and J. Boland, *Representation of a finite graph by a set of intervals on the real line*, Fund. Math., 51 (1962), pp. 45–64.

[14] A. L. Liestman and T. Shermer, *Additive graph spanners*, Networks, 23 (1993), pp. 343–364.

[15] M. S. Madanlal, G. Vankatesan, and C. Pandu Rangan, *Tree 3-spanners on interval, permutation and regular bipartite graphs*, Inform. Process. Lett., 59 (1996), pp. 97–102.

[16] T. A. McKee, *personal communication to E. Prisner*, 1995.

[17] D. Peleg and A. A. Schäffer, *Graph spanners*, J. Graph Theory, 13 (1989), pp. 99–116.

[18] D. Peleg and J. D. Ullman, *An optimal synchronizer for the hypercube*, in Proceedings of the 6th ACM Symposium on Principles of Distributed Computing, 1987, pp. 77–85.

[19] E. Prisner, *Distance approximating spanning trees*, in STACS 97, Lecture Notes in Comput. Sci. 1200, Springer-Verlag, Berlin, 1997, pp. 499–510.

[20] J. Soares, *Graph spanners: A survey*, Congr. Numer., 89 (1992), pp. 225–238.

[21] G. Venkatesan, U. Rotics, M. S. Madanlal, J. A. Makowsky, and C. Pandu Rangan, *Restrictions of minimum spanner problems*, Inform. and Comput., 136 (1997), pp. 143–164.

SIAM J. DISCRETE MATH.
Vol. 17, No. 3, pp. 341–360

SPLIT-PERFECT GRAPHS: CHARACTERIZATIONS AND ALGORITHMIC USE[*]

ANDREAS BRANDSTÄDT[†] AND VAN BANG LE[†]

Abstract. Two graphs G and H with the same vertex set V are *P_4-isomorphic* if every four vertices $\{a, b, c, d\} \subseteq V$ induce a chordless path (denoted by P_4) in G if and only if they induce a P_4 in H. We call a graph *split-perfect* if it is P_4-isomorphic to a split graph (i.e., a graph being partitionable into a clique and a stable set). This paper characterizes the new class of split-perfect graphs using the concepts of homogeneous sets and p-connected graphs and leads to a linear time recognition algorithm for split-perfect graphs, as well as efficient algorithms for classical optimization problems on split-perfect graphs based on the primeval decomposition of graphs. The optimization results considerably extend previous ones on smaller classes such as P_4-sparse graphs, P_4-lite graphs, P_4-laden graphs, and (7,3)-graphs. Moreover, split-perfect graphs form a new subclass of brittle graphs containing the superbrittle graphs for which a new characterization is obtained leading to linear time recognition.

Key words. perfect graphs, P_4-structure of perfect graphs, graphs with P_4-structure of split graphs, perfectly orderable graphs, brittle graphs, superbrittle graphs, P_4-sparse graphs, P_4-lite graphs, P_4-laden graphs, good characterization, linear time recognition, primeval decomposition tree

AMS subject classifications. 05C17, 05C75, 05C85

DOI. 10.1137/S0895480100367676

1. Introduction.

Graph decomposition is a powerful tool in designing efficient algorithms for basic algorithmic graph problems such as maximum independent set, minimum coloring, and many others. Recently, the modular, the primeval, and the homogeneous decomposition of graphs attracted much attention. The last two types of decomposition were introduced by Jamison and Olariu [42] (see also [5]) and are based on their structure theorem and the concept of P_4-connectedness. A P_4 is an induced path on four vertices. A graph $G = (V, E)$ is *P_4-connected* (*p-connected* for short) if, for every partition V_1, V_2 of V with nonempty V_1, V_2, there is a P_4 of G with vertices in V_1 and in V_2, called *crossing* P_4. It is easy to see that every graph has a unique partition into maximal induced p-connected subgraphs, called *p-connected components* (*p-components* for short), and vertices belonging to no P_4.

We follow this line of research by introducing and characterizing a new class of graphs—the *split-perfect graphs*—for which the p-connected components have a simple structure generalizing split graphs. As usual, a graph is called a *split graph* if its vertex set can be partitioned into a clique and a stable set.

The p-connected components represent the nontrivial leaves in the primeval decomposition tree, and thus some basic algorithmic problems can be solved in linear time along the primeval decomposition tree.

The primeval tree is a generalization of the *cotree* representing the structure of the well-known *cographs*, i.e., the graphs containing no induced P_4. A cograph or its complement is disconnected, and the cotree expresses this in terms of corresponding

[*] Received by the editors February 10, 2000; accepted for publication (in revised form) June 25, 2003; published electronically January 6, 2004. An extended abstract of this paper appeared in *Graph-Theoretic Concepts in Computer Science*, Lecture Notes in Comput. Sci. 1928, Springer–Verlag, Berlin, 2000, pp. 71–82.

http://www.siam.org/journals/sidma/17-3/36767.html

[†] Institut für Theoretische Informatik, Fachbereich Informatik, Universität Rostock, 18051 Rostock, Germany (ab@informatik.uni-rostock.de, le@informatik.uni-rostock.de).

cojoin and join operations. The cotree representation of a cograph is essential in solving various NP-hard problems efficiently for these graphs; see [19, 20] for more information on P_4-free graphs.

The study of P_4-free graphs has motivated considering graphs with few P_4's, such as *P_4-reducible graphs* [37, 40] (no vertex belongs to more than one P_4), *P_4-sparse graphs* [32, 33, 39, 41, 44] (no set of five vertices induces more than one P_4), *P_4-lite graphs* [38] (every set of at most six vertices induces at most two P_4's or a "spider"), and *P_4-laden graphs* [28] (every set of at most six vertices induces at most two P_4's or a split graph). Note that in this order, every graph class mentioned in this paragraph is a subclass of the next one.

Recently, Babel and Olariu [4] considered graphs in which no set of at most q vertices induces more than t P_4's, called *(q, t)-graphs*. The most interesting case is $t = q - 4$: (4,0)-graphs are exactly the P_4-free graphs, (5,1)-graphs are exactly the P_4-sparse graphs, and it turns out that P_4-lite graphs form a subclass of (7,3)-graphs. For all these graphs, nice structural results have been obtained that yield efficient solutions for classical NP-hard problems. Our new class of split-perfect graphs extends all of them.

Another motivation for studying graph classes with special P_4-structure stems from the *greedy coloring heuristic*: Define a linear order $<$ on the vertex set, and then always color the vertices along this order with the smallest available color. Chvátal [17] called $<$ a *perfect order* of G if, for each induced subgraph H of G, the greedy heuristic colors H optimally. Graphs having a perfect order are called *perfectly orderable* (see [34] for a comprehensive survey); they are NP-hard to recognize [46]. Because of the importance of perfectly orderable graphs, however, it is natural to study subclasses of such graphs which can be recognized efficiently. Such a class was suggested by Chvátal in [16]; he called a graph G *brittle* if each induced subgraph H of G contains a vertex that is not an endpoint of any P_4 in H or not a midpoint of any P_4 in H. Brittle graphs are discussed in [35, 50, 51]. Babel and Olariu [4] proved that (7,3)-graphs are brittle, and Giakoumakis [28] proved that P_4-laden graphs are brittle. A natural subclass of brittle graphs, called *superbrittle*, consists of those graphs G in which *every* vertex is not an endpoint of any P_4 in G or not a midpoint of any P_4 in G. Split graphs are superbrittle since in a split graph with clique C and stable set S, every midpoint of a P_4 is in C and every endpoint of a P_4 is in S. Superbrittle graphs are characterized in terms of forbidden induced subgraphs in [47]. We will show that our new class of split-perfect graphs is a subclass of brittle graphs, containing all superbrittle graphs. Moreover, we construct a perfect order of a split-perfect graph efficiently, and we obtain a new characterization of superbrittle graphs leading to a linear time recognition.

Yet another motivation for studying split-perfect graphs stems from the theory of perfect graphs. A graph G is called *perfect* if, for each induced subgraph H of G, the chromatic number of H equals the maximum number of pairwise adjacent vertices in H. For example, all the above-mentioned graphs are perfect. For more information on perfect graphs, see [7, 12, 29]. Recognizing perfect graphs in polynomial time is a major open problem in algorithmic graph theory.[1] Two graphs G and H with the same vertex set V are *P_4-isomorphic* if, for all subsets $S \subseteq V$, S induces a P_4 in G if and only if S induces a P_4 in H. Chvátal [18] conjectured and Reed [48] proved that two P_4-isomorphic graphs are both perfect or both imperfect. Thus, to recognize

[1] Very recently, Chudnovsky et al. [14], Chudnovsky and Seymour [15], and Cornuéjols, Liu, and Vušković [22] have announced that perfect graphs can be recognized in polynomial time.

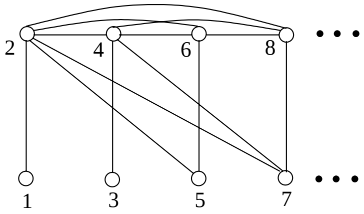


FIG. 1.1. *Elementary graphs illustrated.*

perfect graphs it is enough to recognize the P_4-structure of perfect graphs: given a 4-uniform hypergraph $\mathcal{H} = (V, \mathcal{E})$. Is there a perfect graph $G = (V, E)$ such that $S \in \mathcal{E}$ if and only if S induces a P_4 in G? This was done for the case when the perfect graph G is a tree [25, 10, 11], a block graph [8], the line graph of a bipartite graph [52], a claw-free graph [3], or a bipartite graph [2]. Note that the P_4-structure of a (not necessarily perfect) graph can be recognized in polynomial time [31].

Another question arising from Reed's theorem is the following: Which (perfect) graphs are P_4-isomorphic to a member of a given class of perfect graphs? Let $\mathcal{C}$ be a class of perfect graphs. Graphs P_4-isomorphic to a member in $\mathcal{C}$ are called *$\mathcal{C}$-perfect graphs.* By Reed's theorem, $\mathcal{C}$-perfect graphs are perfect. Moreover, they form a class of graphs which is closed under complementation and contains $\mathcal{C}$ as a subclass. Thus, it is interesting to ask the following question: Assuming that there is a polynomial time algorithm for testing membership in $\mathcal{C}$, can $\mathcal{C}$-perfect graphs be recognized in polynomial time, too? First results in this direction are good characterizations of tree-perfect graphs, forest-perfect graphs [9], and bipartite-perfect graphs [43]. This paper will give a good characterization of split-perfect graphs.

DEFINITION 1.1. *A graph is called* split-perfect *if it is P_4-isomorphic to a split graph.*

Trivial examples of split-perfect graphs are split graphs and P_4-free graphs. Nontrivial examples are induced paths $P_n = v_1 v_2 \cdots v_n$ for any integer n. To see this we need some definitions, following [9]. Let $(v_1, \ldots, v_n)$ be a vertex order of a graph G. Then $N_{>i}(v_i)$ denotes the set of all neighbors v_k of v_i with $k > i$. A vertex order $(v_1, \ldots, v_n)$ of G is said to be *elementary* if for all i

$$N_{>i}(v_i) = \begin{cases} \{v_{i+2}, v_{i+3}, \ldots, v_n\} & \text{for even } i, \\ \{v_{i+1}\} & \text{for odd } i. \end{cases}$$

Graphs having elementary orders are split graphs in which the "odd vertices" v_{2k+1} form a stable set and the "even vertices" v_{2k} form a clique. A graph is said to be *elementary* if it has an elementary order (see Figure 1.1). If the elementary graph has at least 4 vertices, then its partition into a clique and a stable set is unique and can be determined using its degree sequence. Thus, as split graphs in general [30], elementary graphs can be recognized in linear time.

Obviously, $P_n = v_1 v_2 \cdots v_n$ is P_4-isomorphic to the elementary graph consisting of the elementary order $(v_1, \ldots, v_n)$. It can be seen that, for $n \geq 7$, this elementary graph is the only split graph (up to "complementation" and "bipartite complementation") that is P_4-isomorphic to P_n. In section 4, we will extend this example to the so-called *double-split graphs.* Double-split graphs play a key role for characterizing split-perfect graphs.

In section 2, we will show that the class of split-perfect graphs contains all P_4-laden graphs and all (7,3)-graphs (hence all P_4-reducible, P_4-sparse, and P_4-lite

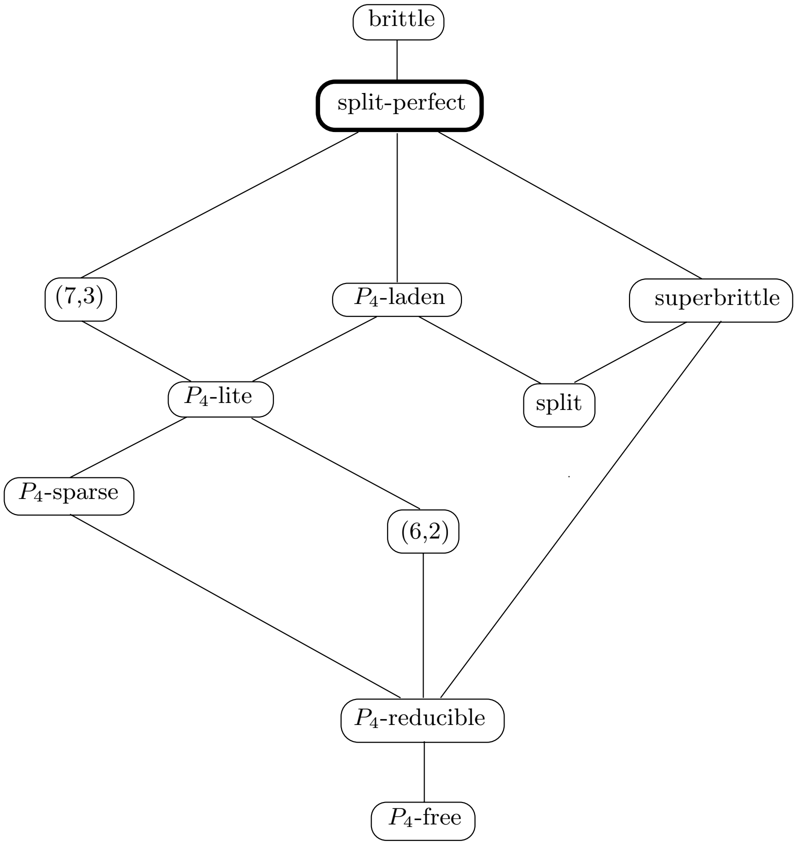


FIG. 1.2. *Relationship between graph classes.*

graphs). The relationship between the above-mentioned graph classes is shown in Figure 1.2.

In section 3, we describe forbidden induced subgraphs of split-perfect graphs, which are needed for characterizing split-perfect graphs.

In section 4, we introduce double-split graphs and show that they are split-perfect. As already mentioned, double-split graphs are of crucial importance for a good characterization of split-perfect graphs.

In section 5, we characterize split-perfect graphs in terms of forbidden subgraphs and in terms of their p-connected components: It turns out that for split-perfect graphs having no homogeneous sets, the p-connected components are double-split graphs or their complements.

In the last section, section 6, we will point out how classical optimization problems such as weighted clique number, weighted chromatic number, weighted independence number, and weighted clique cover number can be solved efficiently, in a divide and conquer manner, on split-perfect graphs using the primeval decomposition tree. These results are based on our good characterization of p-connected split-perfect graphs.

2. Preliminaries. Our notation is quite standard. The neighborhood of the vertex v in a graph G is denoted by $N_G(v)$; if the context is clear, we simply write $N(v)$. The path (respectively, cycle) on m vertices $v_1, v_2, \dots, v_m$ with edges v_iv_{i+1}

(respectively, v_iv_{i+1} and v_1v_m) ($1 \leq i < m$) is denoted by $P_m = v_1v_2 \cdots v_m$ (respectively, $C_m = v_1v_2 \cdots v_mv_1$). The vertices v_1 and v_m are the *endpoints* of the path P_m, and for a P_4 $v_1v_2v_3v_4$, v_2 and v_3 are the *midpoints* of the P_4. Graphs containing no induced subgraphs isomorphic to a graph of a given set H of graphs are called *H-free graphs.* It is well-known that split graphs are exactly the $(C_4, \overline{C_4}, C_5)$-free graphs [26].

For convenience, we often identify sets of vertices of a graph G and the subgraphs induced by these sets in G. Thus, for $S \subseteq V(G)$, S also denotes the subgraph $G[S]$ induced by S.

A set S of at least two vertices of a graph G is called *homogeneous* if $S \neq V(G)$ and every vertex outside S is adjacent to all vertices in S or to no vertex in S. A graph is *prime* if it has at least three vertices and contains no homogeneous set. Obviously, prime graphs and their complements are connected.

A homogeneous set M is *maximal* if no other homogeneous set properly contains M. It is well known that in a connected graph G with connected complement $\overline{G}$, the maximal homogeneous sets are pairwise disjoint (see, e.g., [45]). In this case, the graph G^* obtained from G by contracting every maximal homogeneous set to a single vertex is called the *characteristic graph* of G. Clearly, G^* is prime. We shall use the following useful fact for later discussions (see Figure 3.1 for the graphs G_i).

LEMMA 2.1 (see [36]). *Every prime graph containing an induced C_4 contains an induced $\overline{P_5}$ or G_3 or G_4.*

Throughout this paper, we use the fact that, in a graph $G = (V, E)$, every homogeneous set S contains exactly one vertex of every P_4 crossing S and $V \setminus S$.

For the subsequent structure theorem of Jamison and Olariu we need the following notion: A p-component H of G is called *separable* if it has a partition into nonempty sets H_1, H_2 such that every P_4 with vertices from both H_i's has its midpoints in H_1 and its endpoints in H_2. Note that a p-connected graph is separable if and only if its characteristic graph is a split graph [42].

THEOREM 2.2 (structure theorem [42]). *For an arbitrary graph G, precisely one of the following conditions is satisfied:*

(i) *G is disconnected,*

(ii) *$\overline{G}$ is disconnected,*

(iii) *G is p-connected,*

(iv) *there is a unique proper separable p-component H of G with a partition (H_1, H_2) such that every vertex outside H is adjacent to all vertices in H_1 and nonadjacent to all vertices in H_2.*

Based on this theorem, Jamison and Olariu define the primeval decomposition, which can be described by the primeval decomposition tree and leads to efficient algorithms for a variety of problems if the p-connected components are sufficiently simple. We will show that this is the case for split-perfect graphs.

Note that dividing a graph into p-connected components can be done in linear time (see [6]). This fact together with Proposition 2.3 below allows us to restrict our attention to p-connected split-perfect graphs only.

PROPOSITION 2.3. *A graph is split-perfect if and only if each of its p-connected components is split-perfect.*

Proof. The only if part is clear. To prove the if part, let G be a graph such that each p-connected component A_i ($1 \leq i \leq m$) of G is P_4-isomorphic to a split graph B_i. Let W be the set of all vertices of G not belonging to any P_4. We now construct, inductively, a split graph H_m P_4-isomorphic to G as follows.

First, set $H_1 := B_1 \cup W$. If the split graph H_i ($1 \leq i < m$) is already constructed, then H_{i+1} is obtained from H_i and B_{i+1} by joining every vertex in the clique part of H_i and every vertex of B_{i+1} by an edge.

Clearly, H_m is a split graph. Moreover, B_i ($1 \leq i \leq m$) are exactly the p-connected components of H_m. Thus, H_m is P_4-isomorphic to G. □

Observation 2.4. *Let G be split-perfect and let $H = (C_H, S_H, E_H)$ be a split graph P_4-isomorphic to G. Assume that each of the sets $\{a, b, c, u\}$ and $\{a, b, c, v\}$ induces a P_4 in G. Then exactly one of the following conditions holds:*

(i) *a, b, c induce a path P_3 in H, and u and v are both adjacent in H to an endpoint of the path $H[a, b, c]$. In particular, u and v both belong to the stable-part S_H of H.*

(ii) *The statement* (i) *holds in $\overline{H}$ instead of H. In particular, u and v both belong to the clique-part C_H of H.*

Proof. Since a, b, c, and u induce a P_4 in H, $H[a, b, c]$ must be a P_3, or else a $\overline{P_3}$. The rest follows from the fact that H is a split graph. □

Proposition 2.5. *Let G be a p-connected split-perfect graph. Then every homogeneous set of G induces a P_4-free graph.*

Proof. Assume to the contrary, that there is a homogeneous set S in G which contains an induced P_4 $x_1x_2x_3x_4$. As G is p-connected, there is a crossing P_4 P to the partition S and $V(G) - S$. As S is homogeneous, P has exactly one vertex in S. Let a, b, c be the three vertices of P outside S. Since S is homogeneous, each of the sets $\{a, b, c, x_i\}$, $1 \leq i \leq 4$, induces a P_4 in G. Now, by Observation 2.4, if H is an arbitrary split graph P_4-isomorphic to G, then in H, x_1, x_2, x_3, x_4 are pairwise nonadjacent, or else pairwise adjacent. In particular, $H[x_1, x_2, x_3, x_4]$ cannot be a P_4, a contradiction. □

Proposition 2.6. *Let G be a p-connected graph. G is split-perfect if and only if*

(i) *every homogeneous set of G induces a P_4-free graph, and*

(ii) *G^* is split-perfect.*

Proof. The necessity is clear, because of Proposition 2.5 and the fact that G^* is (isomorphic to) an induced subgraph of G. We now prove the sufficiency. Let G^* be P_4-isomorphic to a split graph H. For each vertex v of G^* let M_v be the corresponding maximal homogeneous set in G. Let H' be the graph obtained from H by replacing each vertex v by the complete graph on vertex set M_v (if v belongs to the clique part of H), respectively, be the stable set M_v (otherwise). Clearly, H' is a split graph. Since the sets M_v contain no P_4, G and H' are P_4-isomorphic (extend a P_4-isomorphism between G^* and H to one between G and H' in a natural way). □

Propositions 2.3 and 2.6 allow us to consider only p-connected split-perfect graphs without homogeneous sets.

Recall that P_4-laden graphs are those graphs in which every set of at most six vertices induces at most two P_4's or a split graph.

Corollary 2.7.

(i) *P_4-laden graphs are split-perfect.*

(ii) (7,3)*-graphs are split-perfect.*

Proof. To prove (i), let G be a p-connected P_4-laden graph. Then

every homogeneous set of G consisting of more than two vertices is a stable set,

otherwise, let M be a homogeneous set with at least three vertices a, b, c, where a and b are adjacent. By the p-connectedness, there is a crossing P_4 P for M and $V(G) - M$. As M is homogeneous, $|V(P) \cap M| = 1$. Now, $(V(P) - M) \cup \{a, b, c\}$ consists of exactly six vertices, induces three P_4's, but does not induce a split graph, a contradiction.

Now, it was proved in [28, Theorem 10] that

$$G^* \text{ is a } P_5 \text{ or } \overline{P_5} \text{ or a split graph.}$$

In particular, G^* is split-perfect and (i) follows from Propositions 2.6 and 2.3.

To prove (ii), we first show the following claims; the first one is easy to see; the second one follows from the known inclusions (6,2) $\subset$ P_4-lite $\subset$ P_4-laden and (i).

Claim 1. *Every graph with at most five vertices, different from the C_5, is split-perfect.* □

Claim 2. (6,2)*-graphs are split-perfect.* □

Now, consider a p-connected (7,3)-graph G. We have to show that G is split-perfect. It was shown in [4, Theorem 4.5] that G has at most six vertices. By Claims 1 and 2, we may assume that G has exactly six vertices and exactly three P_4's.

If G has a homogeneous set, G^* is split-perfect by Claim 1 and every homogeneous set has at most three vertices (otherwise, the p-connectedness would imply that G has four P_4's). Hence, by Proposition 2.6, G is split-perfect. So, let G have no homogeneous set.

If G or $\overline{G}$ has a P_5, say G, then (by considering the neighbors of the vertex outside the P_5) G is a P_6 or the graph with vertices v_i $(1 \leq i \leq 6)$ and edges v_iv_{i+1} $(1 \leq i \leq 5)$, v_2v_6, and v_3v_6 (otherwise G has a homogeneous set or four P_4's). In each case, G is split-perfect.

If G is $(P_5, \overline{P_5})$-free, then G cannot contain an induced C_4 or $\overline{C_4}$. Otherwise, by Lemma 2.1, G would contain a G_3, $\overline{G_3}$, G_4, or $\overline{G_4}$, but each of these graphs has more than three P_4's, a contradiction. Thus, G is $(C_4, \overline{C_4}, C_5)$-free; i.e., G is a split graph and (ii) follows. □

3. Forbidden induced subgraphs for split-perfect graphs. As a consequence of Observation 2.4, we give a list of forbidden induced subgraphs of split-perfect graphs: These are the induced cycles C_k of length $k \geq 5$, the graphs G_i $(1 \leq i \leq 8)$ shown in Figure 3.1, and their complements. It turns out (Theorem 5.1) that these forbidden induced graphs characterize prime split-perfect graphs.

We need some notions. Let G and G' be two graphs with the same vertex set. An induced P_4 in G is *bad* if its vertices do not induce a P_4 in G' (thus, P_4-isomorphic graphs do not have bad P_4's).

Another useful notion is suggested by Observation 2.4: Let G be a split-perfect graph and H a corresponding split graph having the same P_4-structure. We call the clique and the stable set of H the two *classes* of H. Two vertices x, y in G are called *equivalent* $(x \sim y)$ if they are in the same class of H. Clearly, $\sim$ is an equivalence relation on the vertex set of a split-perfect graph.

Now, Observation 2.4 means that in a split-perfect graph G, vertices x and y are in the same class (i.e., $x \sim y$) if there are vertices $a, b, c \in V(G) - \{u, v\}$ such that $\{a, b, c, x\}$ and $\{a, b, c, y\}$ both induce a P_4.

Therefore, in a split-perfect graph, pairwise equivalent vertices induce a P_4-free subgraph.

Recall that a P_4 in a split graph H has its two midpoints in one class and its two endpoints in the other class. Thus, if G is P_4-isomorphic to H, then every P_4 P of G must be *balanced* with respect to H; i.e., P has exactly two vertices in one class and the other two vertices in the other class.

Lemma 3.1. *None of the graphs $C_k, \overline{C_k}$ $(k \geq 5)$, and $G_i, \overline{G_i}$ $(1 \leq i \leq 8)$ in Figure* 3.1 *is split-perfect.*

Proof. Throughout this proof, we will extensively use the facts discussed above.

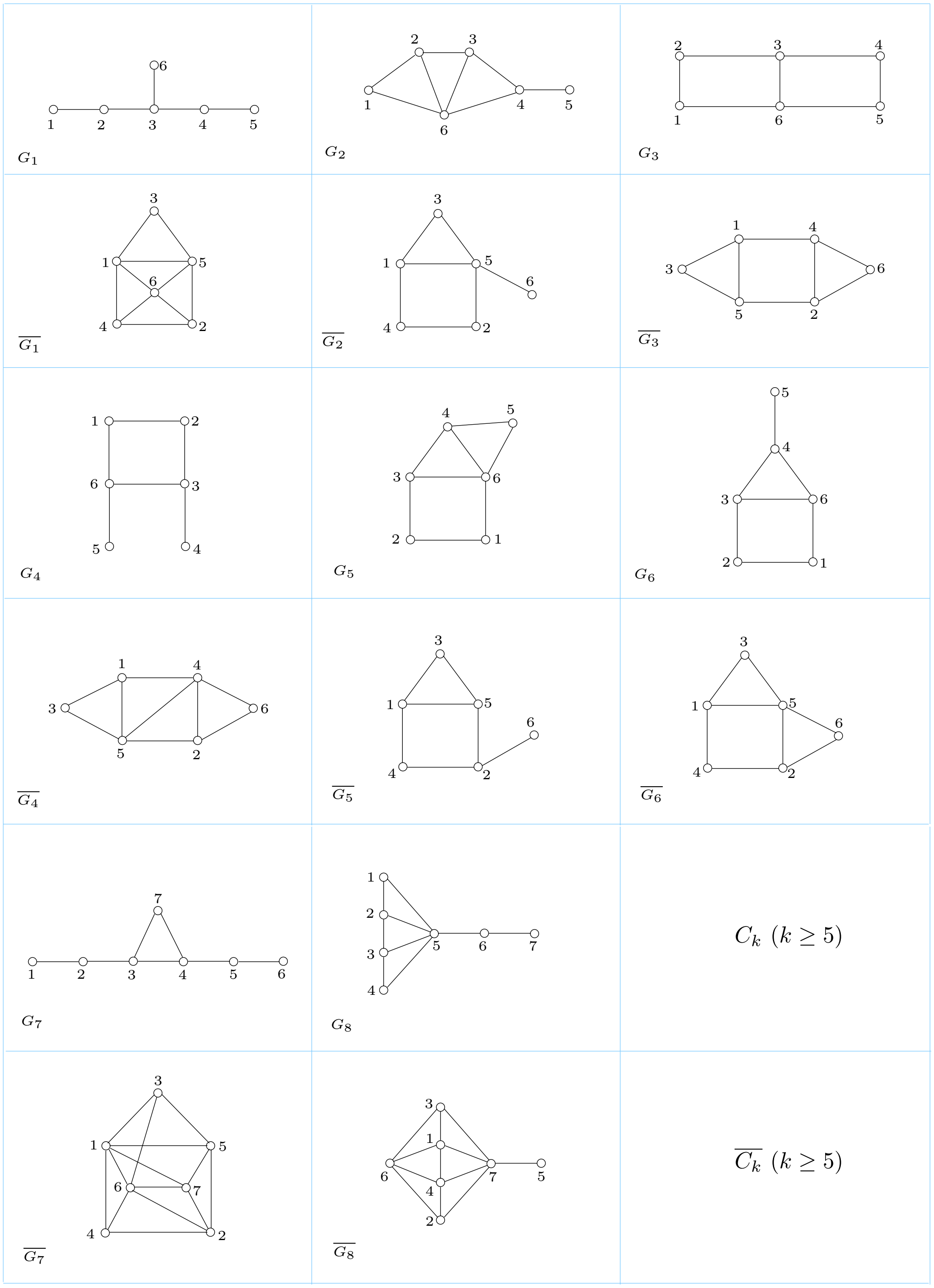


FIG. 3.1. *Forbidden induced subgraphs.*

Note that G is not split-perfect if and only if $\overline{G}$ is not split-perfect. Thus we only show that none of C_k, $k \geq 5$, and G_i, $1 \leq i \leq 8$, is split-perfect.

Consider C_k for odd $k \geq 5$. In this case, all vertices of the C_k are pairwise equivalent, which means that C_k is not split-perfect. (Note that for odd cycles C_k, $k \geq 5$, it also follows from Reed's theorem that they are not split-perfect because they are not perfect.)

Let $k = 2n \geq 6$ and write $C_k = v_1 v_2 \dots v_{2n}$. In this case, all odd vertices v_{2i-1} are pairwise equivalent and all even vertices v_{2i} are pairwise equivalent. Thus, if C_{2n} is split-perfect and H is a corresponding split graph, then, by balance, one class of H consists of exactly the vertices v_{2i-1} and the other class consists of exactly the vertices v_{2i}. Now it is a matter of routine to check that in any realization of the split graph H some P_4 in C_{2n} must be bad.

Assume that $G \in \{G_1, G_3, G_4\}$ is split-perfect and let H be a corresponding split graph. Then $2 \sim 4 \sim 6$. Since the P_4's in G are balanced, the classes of H are $\{2, 4, 6\}$ and $\{1, 3, 5\}$. Again, it is a matter of routine to check that in any realization of the split graph H some P_4 in G must be bad.

Similary, assume that $G \in \{G_2, G_6\}$ is split-perfect and let H be a corresponding split graph. Then $1 \sim 2 \sim 5$. By balance, the classes of H are $\{1, 2, 5\}$ and $\{3, 4, 6\}$. Again, it is a matter of routine to check that in any realization of the split graph H some P_4 in G must be bad.

If G_5 is split-perfect, then 1, 3, 4, 5, and 6 are pairwise equivalent. But then no P_4 in G_5 is balanced.

If G_7 is split-perfect, then $3 \sim 4 \sim 7$. Since every P_4 of G_7 has two vertices in $\{3, 4, 7\}$, it follows by balance that every corresponding split graph H has classes $\{3, 4, 7\}$ and $\{1, 2, 5, 6\}$. Again, it is a matter of routine to check that in any realization of the split graph H some P_4 in G_7 must be bad.

Finally, if G_8 is split-perfect, then 1, 2, 3, and 4 are pairwise equivalent, but induce a P_4. $\square$

4. Double-split graphs.

We define now the class of double-split graphs generalizing the split graphs and playing a key role in the subsequent characterization of split-perfect graphs. As an important step towards this characterization, we will show that double-split graphs are split-perfect.

DEFINITION 4.1. *A graph is called* double-split *if it can be obtained from two disjoint (possibly empty) split graphs $G_L = (Q_L, S_L, E_L)$, $G_R = (Q_R, S_R, E_R)$ and an induced path $P = P[x_L, x_R]$, possibly empty, by adding all edges between x_L and vertices in Q_L and all edges between x_R and vertices in Q_R (see Figure* 4.1*).*

Remark. Every split graph is double-split as the case of an empty path P and an empty split graph G_R shows.

LEMMA 4.2. *Double-split graphs are split-perfect.*

Proof. Let G be a double-split graph consisting of two split graphs $G_L = (Q_L, S_L, E_L)$, $G_R = (Q_R, S_R, E_R)$ with cliques Q_L, Q_R and stable sets S_L, S_R. If the path P connecting G_L and G_R is empty, then G is P_4-isomorphic to the following split graph $H = (Q_L \cup Q_R, S_L \cup S_R, E_H)$ obtained from G_L and G_R by adding a join between Q_L and Q_R and between S_L and Q_R.

Now assume that $P = v_3 v_4 \dots v_i$, $i \geq 3$, such that $x_L = v_3$ is adjacent to all vertices of Q_L and $x_R = v_i$ is adjacent to all vertices of Q_R. We construct a split graph $H = (Q_H, S_H, E_H)$ with the same P_4-structure as G. Hereby we use the fact that induced paths $P' = v_1 v_2 v_3 v_4 \dots v_i v_{i+1} v_{i+2}$ are split-perfect and can be realized by the elementary split graph $G_{P'} = (\{v_2, v_4, v_6, \dots\}, \{v_1, v_3, v_5, \dots\}, E_{P'})$. We will

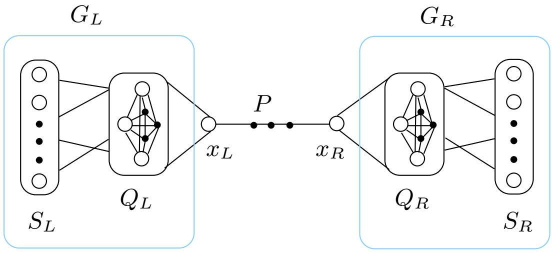


FIG. 4.1. *Double-split graphs illustrated.*

see that this split graph $G_{P'}$ can be extended to H by replacing v_1 by S_L, v_2 by Q_L, v_{i+1} by Q_R, and v_{i+2} by S_R in a suitable way. Moreover, we use the following simple property of split graphs.

CLAIM. *Let $G = (Q, S, E)$ be a split graph and let $G' = (Q, S, E')$ be the following* bipartite complement *of G: For all $x \in Q$ and all $y \in S$, $xy \in E' \Longleftrightarrow xy \notin E$. Then G and G' are P_4-isomorphic.* □

We construct the split graph $H = (Q_H, S_H, E_H)$ depending on the parity of $|P|$; see Figure 4.2.

$$Q_H := \begin{cases} Q_L \cup \{v_4, v_6, \dots, v_{i-1}\} \cup Q_R & \text{if } i \text{ is odd,} \\ Q_L \cup \{v_4, v_6, \dots, v_i\} \cup S_R & \text{otherwise,} \end{cases}$$

$$S_H := \begin{cases} S_L \cup \{v_3, v_5, \dots, v_i\} \cup S_R & \text{if } i \text{ is odd,} \\ S_L \cup \{v_3, v_5, \dots, v_{i-1}\} \cup Q_R & \text{otherwise.} \end{cases}$$

Now E_H consists of the following edges based on the edge set of $G_{P'}$ and on E_L, E_R and depending on the parity of $|P|$:

(1) vertices in Q_H are pairwise adjacent;

(2) the E_H-edge set between S_L and Q_L is E_L;

(3) the E_H-edge set between Q_R and S_R is the bipartite complement of E_R if i is odd and is E_R otherwise;

(4) there is a join between Q_L and S_R (due to the fact that there is an edge between v_2 and v_{i+2} in $G_{P'}$) if i is odd and there is a join between Q_L and Q_R otherwise;

(5) vertices from $\{v_3, v_4, \dots, v_i\}$ have a join to a set from S_L, Q_L, Q_R, S_R if and only if there is an edge in $G_{P'}$ to the corresponding vertex from $\{v_1, v_2, v_{i+1}, v_{i+2}\}$. Thus, for odd i, Q_L has a join to $v_5, v_7, \dots, v_i$, all vertices $x \in \{v_4, v_6, \dots, v_{i-1}\}$ have a join to S_R, and Q_R has a join to v_i; if i is even, then Q_L has a join to $v_5, v_7, \dots, v_{i-1}$, and all vertices $x \in \{v_4, v_6, \dots, v_{i-2}\}$ have a join to Q_R;

(6) the edges between vertices from $v_3, v_4, \dots, v_i$ are the same as in $G_{P'}$.

We claim that G and H are P_4-isomorphic. First we show that every P_4 of G is a P_4 in H. There are the following types of P_4's in G:

(a) P_4's in G_L and P_4's in G_R;

(b) xyv_3v_4 with $x \in S_L$, $y \in Q_L$, $xy \in E_L$ (for $i = 3$ replace v_4 by a vertex $z \in Q_R$);

(c) $xv_3v_4v_5$ for $x \in Q_L$ (for $i = 3$ replace v_4 by a vertex $y \in Q_R$ and v_5 by a vertex $z \in S_R$);

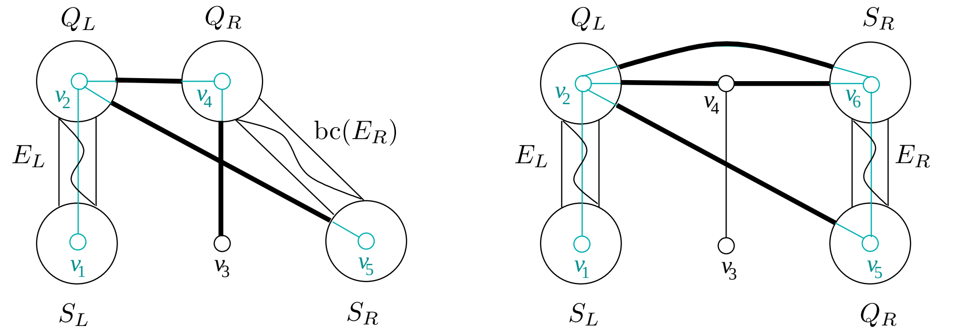


FIG. 4.2. *Construction for $i = 3$ (left) and $i = 4$ (right);* bc(E) *means the bipartite complement of E.*

(d) P_4's in $v_3, v_4, \ldots, v_i$ (for $i \in \{3, 4, 5\}$ there are no such P_4's);

(e) $v_{i-2}v_{i-1}v_ix$ with $x \in Q_R$ (for $i = 3$ this corresponds to case (b), for $i = 4$ replace v_{i-2} by $z \in Q_L$);

(f) $v_{i-1}v_ixy$ with $x \in Q_R$, $y \in S_R$, $xy \in E_R$ (for $i = 3$ replace v_{i-1} by $z \in Q_L$).

Type (a) for G_L is obviously fulfilled by construction of H, and for G_R, the bipartite complement of G_R in H ensures the property if i is odd, and is obvious in the other case.

Types (b), (c), (d), (e) are obviously fulfilled.

Type (f): For the P_4 $v_{i-1}v_ixy$ with $x \in Q_R$, $y \in S_R$, $xy \in E_R$, if i is odd, then xy is not an edge in the bipartite complement of G_R, and thus $v_ixv_{i-1}y$ is a P_4 in H. If i is even, xy is an edge in E_H and $v_{i-1}v_iyx$ is a P_4 in H.

Now consider a P_4 in H. According to the definition of H this is either a P_4 between Q_L and S_L which is the same as in G_L, or a P_4 between Q_R and S_R which, for odd i, is the same as in G_R due to the bipartite complement and, for even i, is obviously the same as in G_R, or a P_4 which goes back to $G_{P'}$ but $G_{P'}$ realizes exactly the P_4's of the induced path P' which are P_4's in G as well. □

Double-split graphs and their complements can be recognized in linear time due to their simple structure as we will show in the appendix.

5. The structure of split-perfect graphs. Now we are able to describe prime split-perfect graphs as follows.

THEOREM 5.1. *Let G be a prime graph. Then the following statements are equivalent:*

(i) *G is split-perfect;*

(ii) *G has no induced subgraphs $C_k, \overline{C_k}$ $(k \geq 5)$, $G_i, \overline{G_i}$ $(1 \leq i \leq 8)$;*

(iii) *G or $\overline{G}$ is a double-split graph.*

Theorem 5.1 and Propositions 2.3 and 2.6 immediately yield the following theorem.

THEOREM 5.2. *A graph G is split-perfect if and only if each of its p-connected components H has the following properties: Every homogeneous set in H induces a P_4-free graph, and H^* is a double-split graph or the complement of a double-split graph.* □

Proof of Theorem 5.1. The implication (i) ⇒ (ii) follows from Lemma 3.1, and the implication (iii) ⇒ (i) follows from Lemma 4.2. Note that these two implications hold in general, not only for p-connected graphs or prime graphs.

We now complete the proof by showing (ii) ⇒ (iii), where we will make use of the primality as follows.

OBSERVATION 5.3. *Let G be prime and let H be a P_4-free induced subgraph of G. If H is not a stable set (a clique, respectively), then there exist adjacent (nonadjacent, respectively) vertices x, y in H and a vertex z outside H such that z is adjacent to x and nonadjacent to y.*

Proof. Assume that H is not a stable set (the case that H is not a clique can be seen similarly). Let $S \subseteq H$ be maximal such that $H[S]$ has no isolated vertices. As H is not a stable set, $|S| \geq 2$. It is well known that P_4-free graphs with at least two vertices contain two vertices u and v with $N(u) = N(v)$ or $N(u) \cup \{u\} = N(v) \cup \{v\}$ (so-called *twins*). Let $\{u, v\}$ be twins in S. As G is prime, there is a vertex $z \notin S$ adjacent to u and nonadjacent to v. By definition of S, $z \notin H$. If u and v are adjacent, then we are done by setting $x = u$ and $y = v$. Thus, let u and v be nonadjacent. By definition of S, u is adjacent to another vertex w in S which is also adjacent to v because $\{u, v\}$ is homogeneous in S. Now, we are done by setting $x = u$, $y = w$ (if z is nonadjacent to w), or $x = w$, $y = v$ (otherwise). ▯

Let G be a prime graph satisfying the statement (ii). If G is $(P_5, \overline{P_5})$-free, then by Lemma 2.1 G cannot contain a C_4 or a $\overline{C_4}$ (otherwise G would contain a G_3, $\overline{G_3}$, G_4, or $\overline{G_4}$). Hence G is $(C_4, \overline{C_4}, C_5)$-free, i.e., G is a split graph and we get (iii).

Therefore, we may assume that G contains a P_5 or a $\overline{P_5}$. By considering complementation if necessary, assume that G has an induced P_5. Consider a longest induced path $P = v_1 v_2 \dots v_k$ in G. By assumption, $k \geq 5$. Now we are going to show, by a number of claims, that G is a double-split graph.

CLAIM NO-MIDDLE. *For every $2 < i < k-1$,*

$$(N(v_{i-1}) \cap N(v_{i+1})) - (N(v_{i-2}) \cup N(v_{i+2})) = \{v_i\}.$$

Proof. Let $H = (N(v_{i-1}) \cap N(v_{i+1})) - (N(v_{i-2}) \cup N(v_{i+2}))$. Then H induces a P_4-free graph, otherwise G would have a $\overline{G_8}$. Thus, assuming $H \neq \{v_i\}$, H has twins $\{x, y\}$. As G has no homogeneous set, there is a vertex $z \notin H$ such that $zx \in E(G)$ but $zy \notin E(G)$. We distinguish between three cases.

Case 1. z is adjacent to both v_{i-1} and v_{i+1}.

By definition of H and $z \notin H$, z must be adjacent to v_{i-2} or v_{i+2}. By symmetry, let $zv_{i-2} \in E(G)$. Now, if z is also adjacent to v_{i+2}, then $v_{i-2}, v_{i-1}, v_{i+1}, v_{i+2}, y, z$ induce a $\overline{G_6}$. If z is nonadjacent to v_{i+2}, then the same vertices induce a $\overline{G_5}$. Case 1 is settled.

Case 2. z is adjacent to v_{i-1} and nonadjacent to v_{i+1} (or vice versa).

Then z cannot be adjacent to v_{i+2} (otherwise there is a C_5). Now, if x and y are adjacent, then there is a G_2, and if x, y are nonadjacent, then there is a $\overline{G_5}$. Case 2 is settled.

Case 3. z is nonadjacent to both v_{i-1} and v_{i+1}.

First, assume $xy \in E(G)$. Then z cannot be adjacent to v_{i-2} or to v_{i+2} (otherwise there is a G_5). But then $v_{i-2}, v_{i-1}, x, v_{i+1}, v_{i+2}, z$ induce a G_1. Second, assume $xy \notin E(G)$. Then there is a G_3 (if z is adjacent to v_{i-2}) or a G_4 (otherwise). Case 3 is settled. ▯

Let M be the set of all vertices outside P adjacent to a vertex in P but not to all vertices in P.

CLAIM N. *For every $v \in M$, $N(v) \cap P = \{v_2\}$ or $\{v_2, v_3\}$ or $\{v_1, v_2, v_3\}$ or $\{v_{k-1}\}$ or $\{v_{k-2}, v_{k-1}\}$ or $\{v_{k-2}, v_{k-1}, v_k\}$.*

Proof. Since G does not have a C_ℓ ($\ell \geq 5$), G_2, $\overline{G_2}$, G_3, G_5, or $\overline{G_6}$, every vertex in M has at most three neighbors in P. We distinguish between three cases.

Case 1. $|N(v) \cap P| = 3$.

Then $N(v) \cap P$ is a subpath of P, otherwise G would have a C_ℓ for some $\ell \geq 5$, or a G_3 or $\overline{G_5}$ or a G_6. Thus $N(v) \cap P = \{v_{i-1}, v_i, v_{i+1}\}$ for some suitable i. Now, by Claim No-Middle, $i = 2$ or $i = k - 1$, and Case 1 is settled.

Case 2. $|N(v) \cap P| = 2$.

We first claim that $N(v) \cap P$ is a subpath of P. Assume to the contrary that the two neighbors of v in P are nonadjacent. Then there is a suitable i such that $N(v) \cap P = \{v_{i-1}, v_{i+1}\}$, otherwise G would have a C_ℓ for some $\ell \geq 5$. Now, by Claim No-Middle, $i = 2$ or $i = k - 1$. By symmetry we only consider the case $N(v) \cap P = \{v_1, v_3\}$.

Let $H = (N(v_1) \cap N(v_3) \cap M) \cup \{v_2\}$. Note that no vertex in H is adjacent to a v_j, $j \geq 4$ (as we have seen in Case 1). Thus H is P_4-free (otherwise G would have a G_8). Since H is not a clique (it contains v and v_2), there exist, by Observation 5.3, nonadjacent vertices $x, y \in H$ and a vertex $z \notin H$ adjacent to x but nonadjacent to y. Note that $z \in M$: If z is nonadjacent to P, then v_1, v_3, x, y, z, v_4 induce a G_4. If z is adjacent to all v_i's, then v_1, v_3, v_4, v_5, y, z induce a G_5. Now, if $zv_3 \in E(G)$, then $zv_1 \notin E(G)$ (otherwise $z \in H$). But then v_1, v_3, v_4, x, y, z induce a $\overline{G_2}$ or a G_5. Thus, $zv_3 \notin E(G)$. But then v_1, v_3, v_4, x, y, z induce a G_3 or contain a C_5 depending on $zv_1 \in E$ (if $zv_4 \in E(G)$) or a G_4 or a $\overline{G_5}$ (if $zv_4 \notin E(G)$).

We have shown that the two neighbors of v on P are v_i and v_{i+1} for some suitable i. Since G has no G_7, $i \in \{1, 2, k-1, k-2\}$. We are going to show that $i \in \{2, k-2\}$ holds. By symmetry, we only show $i \neq 1$.

Assume to the contrary that $i = 1$. Let $H = N(v_2) - N(v_3)$. Then no vertex in H is adjacent to v_j, $j \geq 4$ (as we have seen in Case 1). Thus, H is P_4-free (otherwise G would have a G_8). Since H is not a stable set (it contains v and v_1), there exist, by Observation 5.3, adjacent vertices $x, y \in H$ and vertex $z \notin H$ adjacent to x but nonadjacent to y. If $zv_2 \in E(G)$, then $zv_3 \in E(G)$ (otherwise $z \in H$) and v_2, v_3, v_4, x, y, z induce a G_2 or a $\overline{G_4}$. Thus $zv_2 \notin E(G)$, hence also $zv_3 \notin E(G)$ (otherwise v_2, v_3, v_4, x, y, z induce a $\overline{G_5}$ or a $\overline{G_3}$). But then $zxv_2v_3 \cdots v_k$ is an induced path longer than P, or else z is adjacent to some v_j, $j \geq 4$, yielding a C_{j+1}.

This shows that $i \neq 1$ and, by symmetry, $i \neq k - 1$. We have proved Claim N in Case 2.

Case 3. $|N(v) \cap P| = 1$.

Then $N(v) \cap P = \{v_2\}$ or $N(v) \cap P = \{v_{k-1}\}$. Otherwise G would have a G_1, or there is an induced path longer than P. Claim N is proved in Case 3. □

Let $Q_L = N(v_3) - (N(v_4) \cup N(v_5))$.

CLAIM QL. *Q_L is a clique.*

Proof. First note that Q_L induces a P_4-free graph (otherwise G would have a G_8). Now, assume to the contrary that Q_L is not a clique. By Observation 5.3, there exist nonadjacent vertices $x, y \in Q_L$ and vertex $z \notin Q_L$ adjacent to x but nonadjacent to y. We distinguish between two cases.

Case 1. z and v_3 are nonadjacent.

If $zv_4 \notin E(G)$, then x, y, z, v_3, v_4, v_5 induce a G_1 if $zv_5 \notin E$, else z, x, v_3, v_4, v_5 is a C_5. If $zv_4 \in E(G)$, then, by Claim N, $zv_5 \notin E(G)$ and G has a G_4. Case 1 is settled.

Case 2. z and v_3 are adjacent.

Then $z \in M$. Because, if z is adjacent to all v_i's, then y cannot be adjacent to v_2 (otherwise G would have a $\overline{G_4}$ induced by y, v_2, v_3, v_4, v_5, and z). By Claim N, y is also nonadjacent to v_1. But then G has a G_1.

Now, by definition of Q_L, z must be adjacent to v_4 or v_5, and by Claim N, z is adjacent to v_4 and nonadjacent to v_2. Thus x cannot be adjacent to v_2, otherwise

v_2, v_3, v_4, v_5, x, z induce a G_2 (if $zv_5 \notin E(G)$) or a $\overline{G_4}$ (if $zv_5 \in E(G)$). Therefore, by Claim N, x cannot be adjacent to v_1. But then $v_1, v_2, v_3, v_4, v_5, x$ induce a G_1. Case 2 is settled. □

Let T be the set of all vertices that are adjacent to all vertices in P, and let $S_L = N(Q_L) - (\{v_3\} \cup T)$.

Claim SL. *S_L is a stable set.*

Proof. We first show that

$$v \in S_L \Longrightarrow vv_i \notin E(G), \qquad i \geq 3. \tag{5.1}$$

Proof of (5.1). Assume first that v is adjacent to v_3. By definition of Q_L, v must be adjacent to v_4 or to v_5 (otherwise v would belong to Q_L, contradicting $v \in S_L$). Thus, by Claim N, v is adjacent to v_4 and is nonadjacent to v_1, v_2. Now, a neighbor x in Q_L of v together with v_1, v_2, v_3, v_4, and v induce a G_2 (if $xv_1 \notin E(G)$) or a $\overline{G_4}$ (otherwise). We have shown that v is nonadjacent to v_3. Next, if vv_4 is an edge, then, by Claim N, v is nonadjacent to v_1 and v_2, and so a neighbor x in Q_L together with v_1, v_2, v_3, v_4, v induce a G_6 or a G_5. Thus v is nonadjacent to v_4. Finally, v cannot be adjacent to v_i for any $i \geq 5$ because G does not have a C_ℓ, $\ell \geq 5$. Thus, (5.1) is proved. □

Next, we show that

$$\text{for every two adjacent vertices } u, v \in S_L,\ N(u) \cap Q_L = N(v) \cap Q_L. \tag{5.2}$$

Proof of (5.2). Assume that there is a vertex $x \in Q_L$ adjacent to u but nonadjacent to v, say. Let $y \in Q_L$ be a neighbor of v. Then by (5.1), u, v, x, y, v_3, v_4 induce a G_2 (if $yu \in E(G)$) or a G_6 (otherwise). This contradiction proves (5.2). □

We furthermore show that

$$S_L \text{ induces a } P_4\text{-free graph.} \tag{5.3}$$

Proof of (5.3). If not, then by (5.2), there is a vertex in Q_L adjacent to all vertices of a P_4 in S_L. By (5.2), G would have a G_8. This proves (5.3). □

Now, to finish the proof of Claim SL, assume that S_L is not a stable set. By Observation 5.3, there exist adjacent vertices $u, v \in S_L$ and vertex $w \notin S_L$ adjacent to u but nonadjacent to v. By (5.2), $w \notin Q_L$.

Since $w \notin S_L$, w cannot have a neighbor in Q_L, and it can be seen, as in the proof of (5.1), that w cannot be adjacent to v_i, $i \geq 3$. But then $wuxv_3v_4 \cdots v_k$, where $x \in Q_L$ is a neighbor of u, is an induced path longer than P. The proof of Claim SL is complete. □

Let $Q_R = N(v_{k-2}) - (N(v_{k-3}) \cup N(v_{k-4}))$ and $S_R = N(Q_R) - (\{v_{k-2}\} \cup T)$. By symmetry, we have the following claims.

Claim QR. *Q_R is a clique.* □

Claim SR. *S_R is a stable set.* □

Note that from the definition it follows that $Q_L \cap Q_R = \emptyset$, and from Claim N and the forbidden $\overline{G_6}$ it follows that $S_L \cap S_R = \emptyset$.

Claim NOE (no other edge). *There is no edge between $Q_L \cup S_L$ and $Q_R \cup S_R$.*

Proof. Let $x \in Q_L \cup S_L$ and $y \in Q_R \cup S_R$ be two adjacent vertices. Since P is an induced path and by Claim N, $x, y \notin \{v_1, v_2, v_{k-1}, v_k\}$. Then $x \notin Q_L$ (otherwise y would belong to S_L) and $y \notin Q_R$ (otherwise x would belong to S_R). Thus, $x \in S_L$ and $y \in S_R$, yielding a C_k, $k \geq 5$. This contradiction proves Claim NOE. □

Claim NOV (no other vertex). *$V(G) = P \cup M \cup S_L \cup S_R \cup T$.*

Proof. If there is a vertex $v \notin P \cup M \cup S_L \cup S_R \cup T$, then, as G is connected (it has no homogeneous set), v must be adjacent to some vertex in $S_L \cup S_R$. But then there is an induced path longer than P. □

Claim T. *$T = \emptyset$, i.e., there is no vertex adjacent to all vertices of P.*

Proof. Assume there is a vertex v adjacent to all v_i's. Then v is adjacent to all vertices in Q_L (and in Q_R), otherwise G would have a $\overline{G_4}$. Also, v is adjacent to all vertices in S_L (and in S_R), otherwise G would have a G_2.

Thus, every vertex from T is adjacent to all vertices in $G-T$, implying, by Claim NOV, that $G-T$ is a homogeneous set in G. This contradiction proves Claim T. □

It follows from the claims that G is a double-split graph (with the two split graphs formed by Q_L, S_L and Q_R, S_R, respectively). The proof of Theorem 5.1 is complete. □

Corollary 5.4. *Split-perfect graphs can be recognized in linear time.*

Proof. This follows from Theorem 5.2 and the facts that

- the p-connected components of a graph can be found in linear time [6];
- all maximal homogeneous sets of a (p-connected) graph G can be found in linear time [23, 24, 45];
- P_4-free graphs can be recognized in linear time [21] (for a new and simpler 3-sweep lexicographic breadth-first search algorithm recognizing P_4-free graphs in linear time, see [13]); and
- double-split graphs and their complements can be recognized in linear time (see the appendix). □

In the remainder of this section we will show that the class of split-perfect graphs lies between the classes of superbrittle graphs and of brittle graphs. We first give a new characterization of superbrittle graphs in the following theorem.

Theorem 5.5. *A graph G is superbrittle if and only if for each of its p-connected components H of G,*

(i) *the homogeneous sets of H are cographs, and*

(ii) *the characteristic graph H^* is a split graph.*

Proof. Assume first that G is superbrittle. Then, since the graphs G_8 and $\overline{G_8}$ (see Figure 3.1) are not superbrittle, homogeneous sets in p-connected components are P_4-free; otherwise a crossing P_4 leads to an induced subgraph G_8 or $\overline{G_8}$. Now we show condition (ii). Note first that obviously superbrittle graphs are also $(P_5, \overline{P_5}, C_5, G_4, \overline{G_4})$-free (for G_4 and $\overline{G_4}$, see Figure 3.1). Then, due to Lemma 2.1, H^* is C_4-free since a C_4 in a characteristic graph extends into a $\overline{P_5}$ or G_3 or G_4 but the G_3 contains a P_5. The same holds for the complements which means that H^* and its complement are chordal, i.e., H^* is a split graph.

Now let G be a graph fulfilling the conditions (i) and (ii) for all its p-connected components. We are going to show that G is superbrittle. Since the property to be superbrittle is a P_4 condition, it is sufficient to show that the p-connected components H of G are superbrittle. Note that split graphs are superbrittle, i.e., H^* is superbrittle. Furthermore, by substituting cographs as homogeneous sets into vertices of a split graph, no midpoint of a P_4 in H^* can become an endpoint in H and no endpoint of a P_4 in H^* can become a midpoint in H since homogeneous sets contain at most one vertex of a P_4. This shows that H is superbrittle, and thus G is superbrittle. □

Theorem 5.5 immediately implies the following.

Corollary 5.6. *Superbrittle graphs are split-perfect and can be recognized in linear time.*

Corollary 5.7. *Split-perfect graphs are brittle. Moreover, a perfect order of a split-perfect graph can be constructed efficiently.*

Proof. Since there is no crossing P_4 for two p-connected components, a graph is brittle if and only if each of its p-connected components is brittle. Now, if G is a p-connected split-perfect graph, then G^* is chordal or the complement of a chordal graph (Theorem 5.2); hence G^* is brittle. Let v be a vertex in G^* that is not an endpoint (a midpoint) of any P_4 in G^*. Then, by Proposition 2.5, every vertex in the homogeneous set in G corresponding to v is not an endpoint (a midpoint, respectively) of any P_4 in G. Since every induced subgraph of a split-perfect graph is again split-perfect, it follows that split-perfect graphs are brittle.

Moreover, a perfect order of a split-perfect graph can be constructed as follows: Note that a perfect order of a chordal graph (the complement of a chordal graph) can be found by constructing a perfect elimination order and reversing its order. Now, a perfect order of G^* yields, in a natural way, a perfect order of G. Combining these perfect orders on the p-connected components in an arbitrary sequence, we obtain a perfect order of a split-perfect graph. $\square$

6. Optimization in split-perfect graphs. As already mentioned, Theorem 2.2 implies a decomposition scheme, called *primeval decomposition*, for arbitrary graphs. The corresponding tree representation, called *primeval tree*, has the p-connected components and vertices not belonging to any P_4 of the considered graph as its leaves.

The important features of the primeval tree of a given graph G are the following:

- If an optimization problem such as weighted clique number, weighted chromatic number, weighted independence number, and weighted clique cover number can be solved efficiently on the p-connected components of G, then one can also efficiently solve the problem on the whole graph G; see, for example, [1].
- The primeval tree can be constructed in linear time; see [6].

Based on these facts, linear time or at least polynomial time algorithms have been found for classical NP-hard problems on many graph classes such as $(q, q-4)$-graphs and various subclasses. We now point out how to compute the weighted clique size $\omega_{\mathrm{w}}(G)$ and the weighted independence number $\alpha_{\mathrm{w}}(G)$ for p-connected split-perfect graphs G efficiently.

First, we shall use the following facts:

- The weighted clique number of a chordal graph can be computed in linear time (well known).
- The weighted independence number of a chordal graph can be computed in linear time as pointed out by Frank [27].

Second, let H be a homogeneous set in G and let G/H be the graph obtained from G by contracting H to a single vertex v_H. Then it is well known (and easy to see) that

$$\omega_{\mathrm{w}'}(G/H) = \omega_{\mathrm{w}}(G), \quad \text{respectively,} \quad \alpha_{\mathrm{w}'}(G/H) = \alpha_{\mathrm{w}}(G),$$

where the weighting w' is obtained from w by defining $\mathrm{w}'(v_H) = \omega_{\mathrm{w}}(G[H])$, respectively, $\mathrm{w}'(v_H) = \alpha_{\mathrm{w}}(G[H])$.

Thus, if $\omega_{\mathrm{w}}(G^*)$ and $\omega_{\mathrm{w}}(H)$ (respectively, $\alpha_{\mathrm{w}}(G^*)$ and $\alpha_{\mathrm{w}}(H)$), H a homogeneous set in G, can be computed in linear time, then $\omega_{\mathrm{w}}(G)$ (respectively, $\alpha_{\mathrm{w}}(G)$) can be computed in linear time, too.

Now, if G is a p-connected split-perfect graph, then by Theorem 5.1, G^* is a double-split graph or the complement of a double-split graph. In any case, G^* is a chordal graph or the complement of a chordal graph. If G is chordal, then $\omega_{\mathrm{w}}(G^*)$

and $\alpha_{\mathrm{w}}(G^*)$ can be computed in linear time. If G^* is the complement of a chordal graph, then, by considering $\overline{G^*}$, $\omega_{\mathrm{w}}(G^*)$ and $\alpha_{\mathrm{w}}(G^*)$ can be computed in $O(n^2)$ time (n is the vertex number of G). Furthermore, by Proposition 2.5, every homogeneous set H of G induces a P_4-free graph; hence $\omega_{\mathrm{w}}(H)$ and $\alpha_{\mathrm{w}}(H)$ can be computed in linear time. This and the facts that the primeval tree of G as well as all maximal homogeneous sets of G can be found in linear time show that $\omega_{\mathrm{w}}(G)$ and $\alpha_{\mathrm{w}}(G)$ can be computed in $O(n^2)$ time.

The problems of weighted chromatic number and weighted clique cover number can be solved similarly; we omit the details. Note that for perfect graphs in general and in particular for split-perfect graphs, the weighted chromatic number equals the weighted clique number, and the weighted independence number equals the weighted clique cover number. Thus, we can state the following result.

Theorem 6.1. *The weighted clique number, the weighted chromatic number, the weighted independence number, and the weighted clique cover number of a split-perfect graph can be computed in $O(n^2)$ time.*

Appendix. Linear-time recognition of double-split graphs and their complements. Let $DS(k)$ denote the class of double-split graphs (H_1, P, H_2) with split graphs H_1 and H_2 and k vertices in the induced path P connecting H_1 with H_2, and let $DS = \bigcup_{k\geq 1} DS(k)$.

Theorem A.1. *Double-split graphs and their complements can be recognized in linear time.*

Proof. For a given graph $G = (V, E)$ we have to check whether there is a $k \geq 1$ such that $G \in DS(k)$. Observe that for $G = (H_1, P, H_2) \in DS(k)$ with $k \geq 3$, the path $P = x_1 \ldots x_k$ contains at least one inner vertex of degree 2.

Thus, in order to check whether $G \in DS(k)$ for $k \geq 3$, determine the set D_2 of vertices of degree 2 in G (in the nondegenerate case, D_2 contains no clique vertices from H_1, H_2 and thus D_2 is stable) and check whether $G \setminus D_2$ is the disjoint union of two split graphs H_1', H_2'. Moreover, check whether D_2 is the disjoint union of an induced path P' (the inner vertices of P) and a stable set S'. S_i' consists of the vertices in S' adjacent to some vertex in H_i' for $i \in \{1, 2\}$ (i.e., $H_i' \cup S_i'$ is a split graph H_i with the property that the left (right) endvertex of P' is adjacent to exactly one clique vertex of H_1 (H_2, respectively)).

Now consider the case $G \in DS(1)$ or $G \in DS(2)$. We give an argument using P_4 properties that is similar for the complement graphs.

Case ($G \in DS(1)$). For a given G we have to identify the vertex x_1 of P. If $G \in DS(1)$, G has the following two types of P_4's:

(1) P_4's *abcd* contained in H_1 (H_2, respectively);

(2) P_4's abx_1d containing x_1 as a midpoint.

Thus for a given G, find a P_4 in linear time if there is any (the case that G contains no P_4 reduces to threshold graphs or two cliques intersecting in exactly one vertex), and check whether one of the midpoints of the P_4 (of type (2)) is a cutpoint of G such that the connected components are split graphs and the midpoint is completely adjacent to both of the cliques. If none of the midpoints is a cutpoint, then check the P_4 *abcd* (of type (1)) for the following property: Let $N := N(b) \cap N(c) \cap \overline{N}(a) \cap \overline{N}(d)$, where $\overline{N}(v)$ is the set of all nonneighbors of v. Check whether the two nontrivial connected components of $G' := G \setminus N$ are split graphs. If yes, then one of these split graphs (namely the one not containing the P_4) must have exactly one neighbor x_1 in N. Now check whether the neighborhoods of x_1 in the two components H_1, H_2 of $G \setminus \{x_1\}$ are cliques C_1, C_2 such that $H_i \setminus C_i$ are stable.

Case ($G \in DS(2)$). For a given G we have to identify the vertices x_1, x_2 of P. If $G \in DS(2)$, G has the following three types of P_4's:

(1) P_4's $abcd$ contained in H_1 (H_2, respectively);

(2) P_4's abx_1x_2 containing x_1 as a midpoint and x_2 as an endpoint;

(3) P_4's ax_1x_2b containing x_1, x_2 as midpoints.

Again we start with determining any P_4 in G. For types (2) and (3), try determining whether the midpoints of the P_4 are cutpoints and the connected components fulfill the required properties. For type (1), similar arguments as in case $G \in DS(1)$ will work.

Let co-$DS(k)$ denote the complement graphs of $DS(k)$ graphs. We first describe linear time recognition of co-$DS(k)$ graphs for $k \geq 3$. As for $DS(k)$ graphs, the inner vertices of the path P have to fulfill a degree condition which is now degree $n - 3$. Thus, in order to check whether $G \in$ co-$DS(k)$ for $k \geq 3$, determine the set D_{n-3} of vertices of degree $n - 3$ in G and check whether $G \setminus D_{n-3}$ is the join of two split graphs H'_1, H'_2. In order to check this in linear time, use the techniques of [24] in order to determine the (two nontrivial) connected components H'_1, H'_2 in the complement graph $\overline{G}$ for a given G and check whether they are split graphs. Moreover, check whether the connected components of D_{n-3} in the complement graph are an induced path P' (the inner vertices of P) and two sets S'_1, S'_2 such that $H'_i \cup S'_i$ is a split graph for $i \in \{1, 2\}$ with the property that the left (right) endvertex of P' is nonadjacent to exactly one clique vertex of H_1 (H_2, respectively).

Now consider the case $G \in$ co-$DS(1)$ or $G \in$ co-$DS(2)$. In these cases, using P_4 properties, we find the special vertex x_1 (special vertices x_1, x_2, respectively) as for $G \in DS(1)$ or $G \in DS(2)$, and using the techniques of [24], we find the connected components of $\overline{G}$ in linear time on input G. □

Acknowledgment. We are grateful to two anonymous referees for their careful reading and helpful comments.

REFERENCES

[1] L. Babel, *On the P_4-Structure of Graphs*, Habilitationsschrift, TU München, Munich, 1997.

[2] L. Babel, A. Brandstädt, and V. B. Le, *Recognizing the P_4-structure of bipartite graphs*, Discrete Appl. Math., 93 (1999), pp. 157–168.

[3] L. Babel, A. Brandstädt, and V. B. Le, *Recognizing the P_4-structure of claw-free graphs and of a larger class of graphs*, Discrete Math. Theor. Comput. Sci., 5 (2002), pp. 127–146; also available online from http://dmtcs.loria.fr/volumes/abstracts/dm050109.abs.html.

[4] L. Babel and S. Olariu, *On the structure of graphs with few P_4*, Discrete Appl. Math., 84 (1998), pp. 1–13.

[5] L. Babel and S. Olariu, *On the p-connectedness of graphs: A survey*, Discrete Appl. Math., 95 (1999), pp. 11–33.

[6] S. Baumann, *A Linear Algorithm for the Homogeneous Decomposition of Graphs*, Report M–9615, Zentrum Mathematik, TU München, Munich, 1996.

[7] C. Berge and V. Chvátal, eds., *Topics on Perfect Graphs*, Ann. Discrete Math. 21, North–Holland, Amsterdam, 1984.

[8] A. Brandstädt and V. B. Le, *Recognizing the P_4-structure of block graphs*, Discrete Appl. Math., 99 (2000), pp. 349–366.

[9] A. Brandstädt and V. B. Le, *Tree- and forest-perfect graphs*, Discrete Appl. Math., 95 (1999), pp. 141–162.

[10] A. Brandstädt, V. B. Le, and S. Olariu, *Linear-Time Recognition of the P_4-Structure of Trees*, Rutcor research report 19-96, Rutgers University, New Brunswick, NJ, 1996.

[11] A. Brandstädt, V. B. Le, and S. Olariu, *Efficiently recognizing the P_4-structure of trees and of bipartite graphs without short cycles*, Graphs Combin., 16 (2000), pp. 381–387.

[12] A. Brandstädt, V. B. Le, and J. Spinrad, *Graph Classes: A Survey*, SIAM Monogr. Discrete Math. Appl. 3, SIAM, Philadelphia, 1999.

[13] A. Bretscher, D. G. Corneil, M. Habib, and Ch. Paul, *A simple linear time LexBFS cograph recognition algorithm*, in Proceedings of the 29th Workshop on Graph-Theoretic Concepts in Computer Science (Elspeet, The Netherlands), Lecture Notes in Comput. Sci. 2880, Springer-Verlag, Berlin, 2003, pp. 119–130.

[14] M. Chudnovsky, G. Cornuéjols, X. Liu, P. Seymour, and K. Vušković, *Cleaning for Bergeness*, manuscript, 2003.

[15] M. Chudnovsky and P. Seymour, *Recognizing Berge Graphs*, manuscript, 2003.

[16] V. Chvátal, *Perfect Graphs Seminar*, McGill University, Montreal, 1983.

[17] V. Chvátal, *Perfectly ordered graphs*, Ann. Discrete Math., 21 (1984), pp. 63–65.

[18] V. Chvátal, *A semi-strong perfect graph conjecture*, Ann. Discrete Math., 21 (1984), pp. 279–280.

[19] D. G. Corneil, H. Lerchs, and L. Stewart-Burlingham, *Complement reducible graphs*, Discrete Appl. Math., 3 (1981), pp. 163–174.

[20] D. G. Corneil, Y. Perl, and L. K. Stewart, *Cographs: Recognition, applications, and algorithms*, Congr. Numer., 43 (1984), pp. 249–258.

[21] D. G. Corneil, Y. Perl, and L. K. Stewart, *A linear recognition algorithm for cographs*, SIAM J. Comput., 14 (1985), pp. 926–934.

[22] G. Cornuéjols, X. Liu, and K. Vušković, *A Polynomial Algorithm for Recognizing Perfect Graphs*, manuscript, 2003.

[23] A. Cournier and M. Habib, *A new linear algorithm for modular decomposition*, in Trees in Algebra and Programming – CAAP '94, Lecture Notes in Comput. Sci. 787, Springer, Berlin, 1994, pp. 68–84.

[24] E. Dahlhaus, J. Gustedt, and R. M. McConnell, *Efficient and practical modular decomposition*, J. Algorithms, 41 (2001), pp. 360–387.

[25] G. Ding, *Recognizing the P_4-structure of a tree*, Graphs Combin., 10 (1994), pp. 323–328.

[26] S. Földes and P. L. Hammer, *Split graphs*, Congr. Numer., 19 (1977), pp. 311–315.

[27] A. Frank, *Some polynomial algorithms for certain graphs and hypergraphs*, Congr. Numer., 15 (1976), pp. 211–226.

[28] V. Giakoumakis, *P_4-laden graphs: A new class of brittle graphs*, Inform. Process. Lett., 60 (1996), pp. 29–36.

[29] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[30] P. L. Hammer and B. Simeone, *The splittance of a graph*, Combinatorica, 1 (1981), pp. 275–284.

[31] R. B. Hayward, S. Hougardy, and B. A. Reed, *Polynomial time recognition of P_4-structure*, in Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 2002, pp. 382–389.

[32] C. T. Hoàng, *A Class of Perfect Graphs*, Ms. Sc. thesis, School of Computer Science, McGill University, Montreal, 1983.

[33] C. T. Hoàng, *Perfect Graphs*, Ph.D. thesis, McGill University, Montreal, 1985.

[34] C. T. Hoàng, *Perfectly orderable graphs: A survey*, in Perfect Graphs, Wiley-Intersci. Ser. Discrete Math. Optim., J. L. Ramirez-Alfonsin and B. A. Reed, eds., Wiley, Chichester, 2001, pp. 139–166.

[35] C. T. Hoàng and N. Khouzam, *On brittle graphs*, J. Graph Theory, 12 (1988), pp. 391–404.

[36] C. T. Hoàng and B. Reed, *Some classes of perfectly orderable graphs*, J. Graph Theory, 13 (1989), pp. 445–463.

[37] B. Jamison and S. Olariu, *P_4-reducible graphs – a class of uniquely tree representable graphs*, Stud. Appl. Math., 81 (1989), pp. 79–87.

[38] B. Jamison and S. Olariu, *A new class of brittle graphs*, Stud. Appl. Math., 81 (1989), pp. 89–92.

[39] B. Jamison and S. Olariu, *A unique tree representation for P_4-sparse graphs*, Discrete Appl. Math., 35 (1992), pp. 115–129.

[40] B. Jamison and S. Olariu, *A linear-time algorithm to recognize P_4-reducible graphs*, Theoret. Comput. Sci., 145 (1995), pp. 329–344.

[41] B. Jamison and S. Olariu, *Linear time optimization algorithms for P_4-sparse graphs*, Discrete Appl. Math., 61 (1995), pp. 155–175.

[42] B. Jamison and S. Olariu, *p-components and the homogeneous decomposition of graphs*, SIAM J. Discrete Math., 8 (1995), pp. 448–463.

[43] V. B. Le, *Bipartite-perfect graphs*, Discrete Appl. Math., 127 (2003), pp. 581–599.

[44] R. Lin and S. Olariu, *A fast parallel algorithm to recognize P_4-sparse graphs*, Discrete Appl. Math., 81 (1998), pp. 191–215.

[45] R. M. McConnell and J. Spinrad, *Modular decomposition and transitive orientation*, Discrete Math., 201 (1999), pp. 189–241.

[46] M. Middendorf and F. Pfeiffer, *On the complexity of recognizing perfectly orderable graphs*, Discrete Math., 80 (1990), pp. 327–333.

[47] M. Preissmann, D. de Werra, and N. V. R. Mahadev, *A note on superbrittle graphs*, Discrete Math., 61 (1986), pp. 259–267.

[48] B. Reed, *A semi-strong perfect graph theorem*, J. Combin. Theory Ser. B, 43 (1987), pp. 223–240.

[49] D. J. Rose, R. E. Tarjan, and G. S. Lueker, *Algorithmic aspects of vertex elimination on graphs*, SIAM J. Comput., 5 (1976), pp. 266–283.

[50] A. Schäffer, *Recognizing brittle graphs: Remarks on a paper of Hoàng and Khouzam*, Discrete Appl. Math., 31 (1991), pp. 29–35.

[51] J. P. Spinrad and J. L. Johnson, *Brittle, Bipolarizable, and P_4-Simplicial Graph Recognition*, manuscript, Department of Computer Science, Vanderbilt University, Nashville, TN, 1999.

[52] S. Sorg, *Die P_4-Struktur von Kantengraphen bipartiter Graphen*, Diploma thesis, Mathematisches Institut der Universität zu Köln, Cologne, 1997.

SIAM J. Discrete Math.
Vol. 17, No. 3, pp. 361–376

COLORING THE MAXIMAL CLIQUES OF GRAPHS[*]

GÁBOR BACSÓ[†], SYLVAIN GRAVIER[‡], ANDRÁS GYÁRFÁS[†],
MYRIAM PREISSMANN[‡], AND ANDRÁS SEBŐ[‡]

Abstract. In this paper we are concerned with the so-called clique-colorations of a graph, that is, colorations of the vertices so that no maximal clique is monochromatic. On one hand, it is known to be NP-complete to decide whether a perfect graph is 2-clique-colorable, or whether a triangle-free graph is 3-clique-colorable; on the other hand, there is no example of a perfect graph where more than three colors would be necessary. We first exhibit some simple recursive methods to clique-color graphs and then relate the chromatic number, the domination number, and the maximum cardinality of a stable set to the clique-chromatic number. We show exact bounds and polynomial algorithms that find the clique-chromatic number for some classes of graphs and prove NP-completeness results for some others, trying to find the boundary between the two. For instance, while it is NP-complete to decide whether a graph of maximum degree 3 is 2-clique-colorable, $K_{1,3}$-free graphs without an odd hole turn out to be always 2-clique-colorable by a polynomial algorithm. Finally, we show that "almost" all perfect graphs are 3-clique-colorable.

Key words. clique-coloring, hypergraph

AMS subject classifications. 15C15, 15C17

DOI. 10.1137/S0895480199359995

1. Introduction.

A *hypergraph* $\mathcal{H}$ is a pair $(V, \mathcal{E})$, where V is the set of *vertices* of $\mathcal{H}$, and $\mathcal{E}$ is a family of nonempty subsets of V called *edges* of $\mathcal{H}$. In this paper graphs are always undirected, that is, they are hypergraphs where every edge has two elements. A *k-coloration* of $\mathcal{H} = (V, \mathcal{E})$ is a mapping $c : V \to \{1, 2, \ldots, k\}$ such that for all $e \in \mathcal{E}$, $|e| \geq 2$, there exist $u, v \in e$ with $c(u) \neq c(v)$. The *chromatic number* $\chi(\mathcal{H})$ of $\mathcal{H}$ is the smallest k for which $\mathcal{H}$ has a k-coloration. In other words, a k-coloration of $\mathcal{H}$ is a partition $\mathcal{P}$ of V into at most k parts such that no edge of cardinality at least 2 is contained in some $P \in \mathcal{P}$.

As usual, $K_{i,j}$ $(i, j \in \mathbb{N})$ denotes the complete bipartite graph with classes of cardinality i and j; K_n is the complete graph on n vertices, and C_n is a graph on n vertices and n edges forming a circuit. The graph $K_{1,3}$ is also called a *claw*, and $K_3 = C_3$ a *triangle*. A *hole* is an induced chordless cycle with at least five vertices. A *cobipartite graph* is the complement of a bipartite graph.

A graph is called *H-free*, where H is an arbitrary fixed graph, if it does not contain H as an induced subgraph.

In this paper we consider hypergraphs arising from graphs: for a given graph $G = (V, E)$, the *clique-hypergraph of G* is defined as $\mathcal{H}(G) = (V, \mathcal{E})$, where $\mathcal{E} = \{K \subseteq V\ : K$ is a maximal clique of G$\}$. (A set $K \subseteq V$ of vertices is a *clique* if $ab \in E$ holds for all distinct $a, b \in K$, and K is a maximal clique if it is not properly contained in

[*] Received by the editors August 9, 1999; accepted for publication (in revised form) June 21, 2003; published electronically January 6, 2004. This work was supported by projects TEMPRA (Région Rhône-Alpes) and Balaton (CNRS and Hungarian Academy of Sciences).
http://www.siam.org/journals/sidma/17-3/35999.html

[†] Combinatorial Computer Science Research Group, Systems and Control Laboratory, Computer and Automation Research Institute, Hungarian Academy of Sciences, P.O. Box 63, Budapest, H-1518 Hungary (bacso@sztaki.hu, gyarfas@sztaki.hu).

[‡] CNRS, Département de Mathématiques Discrètes, Laboratoire LEIBNIZ-IMAG, 46 avenue Félix Viallet, 38031 Grenoble Cedex 1, France (Sylvain.Gravier@imag.fr, Myriam.Preissmann@imag.fr, Andras.Sebo@imag.fr).

any other clique.) A hypergraph $\mathcal{H}$ will be called a *clique-hypergraph* if $\mathcal{H} = \mathcal{H}(G)$ for some graph G defined on the vertices of $\mathcal{H}$.

A k-coloration of $\mathcal{H}(G)$ will also be called a *k-clique-coloration* of G, and the chromatic number of $\mathcal{H}(G)$ the *clique-chromatic number of* G. We hope it will not be confusing to use in parallel the usual terms *k-coloration* and *chromatic number* $\chi(G)$ of G where $c(u) \neq c(v)$ is required for every edge $uv \in E$. As usual, the maximum size of a clique in G is denoted by $\omega = \omega(G)$ and the maximum size of a stable set (a set of vertices not containing any induced edge) by $\alpha = \alpha(G)(= \omega(\bar{G}))$. We will also use the shorthand notations $\kappa := \kappa(G) := \chi(\mathcal{H}(G))$, $\bar{\kappa} := \kappa(\bar{G})$, $\bar{\chi} := \chi(\bar{G})$.

Note that what we call k-clique-coloration here is called strong k-division by Hoàng and McDiarmid in [7]. The main objective of [7] is to find a k-coloration of the hypergraph of *maximum* cliques, which leads for most part to problems of a different nature from those studied here. However, the theorems of [7] on strong k-divisions are related to some of our results, and we will point out the connections that we have understood.

Before explaining some connections between colorations and clique-colorations of graphs, let us show some essential differences concerning combinatorial properties as well as problem complexity.

1. A basic property of graph colorations is that they also provide proper colorations of all the subgraphs of the colored graph. This allows us to define various notions of "critical graphs" and is extensively used in coloring algorithms and proofs. On the contrary, a clique-coloration of G does not necessarily induce clique-colorations of the subgraphs of G; accordingly, the clique-chromatic number is not necessarily smaller for induced subgraphs.

For example, if G is a (nonempty) graph and G' is obtained from G by adding a vertex of full degree, then $\chi(G') = \chi(G) + 1$ while $\kappa(G') = 2$.

However, a k-clique-coloration of a graph can be defined with the k-coloration of a subgraph. This subgraph is not induced by a set of vertices, but arises by deleting edges and vertices of the graph (see after 3 below). Unfortunately a proper way of doing this depends on the clique-coloration itself: deleting or contracting monochromatic edges in a clique coloration does lead to properly colored graphs.

2. The hereditary property of colorations involves advantageous algorithmic behavior as well: one can color the vertices successively by giving to each new vertex a color different from those already assigned to its neighbors (rules can be defined for the order in which the vertices are colored and for the choice of the color). All vertex-colorations, including the optimal ones, can arise in this way.

A simple but very useful modification of this sequential coloring procedure is to combine it with "bichromatic exchanges" (see, for example, [13]). Such natural procedures do not show up for the clique-coloring number even if some sequential procedures will produce some results in what follows.

3. Some of the most basic problems that are completely trivial for coloring become intractable for clique-coloring: the problem of deciding whether a hypergraph given explicitly admits a 2-coloration is known to be NP-complete [11], even for clique-hypergraphs [10]. Furthermore, just to check whether a given set is a color class in some clique-coloration is NP-hard; see section 2.

Clearly, any k-coloration of G is a k-clique-coloration, whence $\kappa \leq \chi$. Typically κ is much smaller than χ. However, *a graph G has a k-clique-coloration if and only if it has a subgraph H such that*

- *for every maximal clique K of G, $|E(H) \cap E(K)| \geq 1$;*
- *H has a k-coloration.*

Indeed, a k-coloration of H can be arbitrarily extended to a k-clique-coloration of G. Conversely, the edges whose two endpoints have different colors in a k-clique-coloration of G define H with the claimed properties.

If G is triangle-free, then of course $\kappa(G) = \chi(G)$. Since the chromatic number of triangle-free graphs is known to be unbounded [17], we get that the same is true for the clique-chromatic number. Let us recall for further use Mycielski's triangle-free graphs with unbounded chromatic number:

- G_2 consists of two adjacent vertices.
- For any $k > 2$, the graph $G_k = (V_k, E_k)$ is defined by the following:
  - $V_k = V_{k-1} \cup S_k \cup \{x_k\}$, where $V_{k-1} = \{v_1, \ldots, v_{n_{k-1}}\}$ and $S_k = \{s_1, \ldots, s_{n_{k-1}}\}$;
  - the subgraph induced by V_{k-1} is isomorphic to G_{k-1}, and the subgraph induced by S_k is a stable set;
  - there exists an edge $s_i v_j$ if and only if there exists an edge $v_i v_j$;
  - x_k is adjacent to all vertices in S_k and to no other vertex.

It is easy to show by induction that G_k is triangle-free and $\chi(G_k) = k$ for all $k \geq 2$. It is also easy to check that $\chi(G_k \setminus \{e\}) = k - 1$ for every edge e of G_k.

The clique-chromatic number is unbounded already for the line-graphs of very particular graphs. Indeed, from the existence of Ramsey numbers we get that for any fixed k there exists $N_k \in \mathbb{N}$ so that for all $n \geq N_k$, every k-edge-coloration of K_n contains a monocolored triangle. A triangle of K_n is a maximal clique in the line-graph L_n of K_n. Therefore $\kappa(L_n) \geq k + 1$ if $n \geq N_k$.

However, in [4] (reported also in [8]), the following question is asked.

Question 1. Does there exist some constant C so that it is always possible to C-color the clique-hypergraph $\mathcal{H}(\mathcal{G})$ of a perfect graph G?

Recall that a graph is *perfect* if, for every induced subgraph G', $\chi(G') = \omega(G')$; that is, the chromatic number of G' is equal to its maximum clique size.

Duffus et al. [4] observe that the answer to Question 1 is positive for two subclasses of perfect graphs: the clique-chromatic number of comparability graphs is at most 2, and that of cocomparability graphs is at most 3 by a result of Duffus, Kierstead, and Trotter [3]. In this paper we show that the answer to Question 1 is yes in some other cases, and again with $C = 2$ or $C = 3$. We do not have any example of a perfect graph, and not even of an odd-hole-free graph, with clique-chromatic number greater than 3.

Let us finally introduce some more notation and terminology. For $U \subseteq V$ we will use the notation $N(U) := \{v \in V : v \notin U$, and there exists $u \in U$ such that $uv \in E\}$, $N[U] := N(U) \cup U$. Instead of $\{x\}$ we will often write x. The *border* $B(U)$ of U is $N(U) \cup N(V \setminus U)$; that is, $B(U)$ is the set of vertices of U or $V \setminus U$ that has a neighbor in $V \setminus U$ or U, respectively. ($B(U) = B(V \setminus U)$). We will say that $u \in U$ is a *border-guard* of U if $N[u] \supseteq B(U)$. Borders and border-guards will be useful for clique-colorations because of the simple fact that any $Q \in \mathcal{E}(\mathcal{H}(G))$ is either entirely contained in U, in $V \setminus U$, or in $B(U)$; in the latter case Q contains all the border-guards of U.

Given $U \subseteq V$ and $u \in U$ it is easy to test whether u is a border-guard of U. This is to be appreciated, because it is not as easy to exhibit a "reasonable" clique-coloration as it is a coloration; the main difficulty is that it is NP-hard already to check whether a given mapping is a clique-coloration! The mentioned properties of border-guards are helpful for achieving these tasks whenever border-guards exist.

In section 2, we analyze various aspects of the complexity of clique-coloring. In section 3, we show some simple but general (greedy) methods to clique-color graphs. In section 4, we exhibit connections between $\kappa(G)$ and other parameters of the graph G. In section 5, we prove that some classes of clique-hypergraphs are 2- or 3-colorable. Finally, in section 6, we show that almost all perfect graphs are 3-clique-colorable.

2. The complexity of clique-coloring.

In this section, we study several aspects of the complexity of clique-coloring.

It is already coNP-complete to check whether a given function c defined on the vertices of a graph is a clique-coloration. More precisely, the following problem is shown to be NP-complete.

MAXIMAL CLIQUE CONTAINMENT.

INPUT: Graph $G = (V, E)$ and $T \subseteq V$.

QUESTION: Is there a maximal clique K of G such that $K \subseteq T$?

Therefore deciding whether a k-clique-coloration exists is not clearly in NP nor clearly in coNP.

THEOREM 1. MAXIMAL CLIQUE CONTAINMENT *is NP-complete and remains NP-complete if the complement of the input graph G is restricted to be $K_{1,4}$-free.*

Proof. The 3-DM (that is, three-dimensional matching; see [5]) can be very simply reduced to this problem (a similar proof of [1] can be shortcut for this simpler situation): let $(X, Y, Z, \mathcal{T})$ be an instance of 3-DM; that is, X, Y, Z are finite sets, $|X| = |Y| = |Z|$, and $\mathcal{T} \subseteq X \cup Y \cup Z$ so that for all $T \in \mathcal{T}$, $|T \cap X| = |T \cap Y| = |T \cap Z| = 1$. Let $\mathcal{E} := \mathcal{T} \cup \{\{y\} : y \in Y\}$.

We let G be the intersection graph of the hypergraph $(X \cup Y \cup Z, \mathcal{E})$, that is, the vertex-set of G is $\mathcal{E}$, and we join two vertices if they intersect. The following statements can be easily checked: $\mathcal{T}$ contains a maximal stable set of G if and only if the 3-DM problem has a solution, that is, if the family $\mathcal{T}$ contains a partition of $X \cup Y \cup Z$; since the cardinality of every set in $\mathcal{E}$ is at most three, G is $K_{1,4}$-free.

Thus the 3-DM problem for $(X, Y, Z, \mathcal{T})$ is reduced to the existence of a maximal clique of $\bar{G}$ contained in $\mathcal{T}$, where $\bar{G}$ is $K_{1,4}$-free. □

If the maximal cliques of a graph are given, it can of course be checked in polynomial time if a coloration is a clique-coloration. So, for general algorithmic considerations it is reasonable to consider the problem in a setting where $\mathcal{H}(G)$ is given as part of the input.

We will in fact consider the following seemingly more general problem.

k-CLIQUE-COLORING.

INPUT: A family $\mathcal{H}$ of maximal cliques of G, and $k \in \mathbb{N}$.

QUESTION: Can $\mathcal{H}$ be k-colored?

The problem of coloring $\mathcal{H}$ is not really more general than that of coloring $\mathcal{H}(G)$. Indeed, adding to G a vertex v_K for every clique $K \in \mathcal{H}(G) \setminus \mathcal{H}$, and joining v_K exactly to the vertices of K, we obtain a graph G' with the property that $\mathcal{H}$ is k-colorable if and only if $\mathcal{H}(G')$ is k-colorable ($k \geq 2$).

This does not mean that $\mathcal{H}$ arises as the hypergraph of *all* the maximal cliques of some graph: let G be the graph consisting of a circuit on 6 vertices and 3 chords forming a triangle T; then $\mathcal{H}(G) \setminus \{T\}$ does not arise as the set of all maximal cliques of a graph.

Notice also that the problem of coloring clique-hypergraphs is more restrictive than that of general hypergraph coloring: the hypergraph $\{1,2\}$, $\{2,3\}$, $\{3,1\}$ does not arise as a clique-hypergraph.

Since the computation of the chromatic number is NP-hard for triangle-free graphs [12], it is also *NP-hard to compute the clique-chromatic number* of triangle-free graphs, even if all the cliques are given explicitly as part of the input.

Quite general classes of hypergraphs can be 2-colored. Using the Lovász local lemma, McDiarmid [15] proves that all hypergraphs whose hyperedges are "large" (in a well-defined sense), as compared to the degrees, are 2-colorable. Almost all perfect graphs are 3-clique-colorable (see section 6), but deciding if a perfect graph of maximum clique-size four is 2-clique-colorable is already NP-complete, by Kratochvíl and Tuza [10]. On the other hand, Mohar and Škrekovski [16] have shown that every planar graph is 3-clique-colorable, and Kratochvíl and Tuza [10] proposed a polynomial algorithm to decide if a planar graph is 2-clique-colorable (the set of cliques is given in the input).

The following result is inspired by the methods of [10].

THEOREM 2. *2-clique coloring is NP-complete even if the input graph G is restricted to be of maximum degree* 3.

Proof. We use the not-all-equal satisfiability problem (NAE-SAT), which is known to be NP-complete [21].

NAE-SAT.

INPUT: A set X of Boolean variables and a collection C of clauses (set of literals over U), each clause containing three different literals.

QUESTION: Is there a truth assignment for X such that every clause contains at least one true and at least one false literal?

Given an instance $\mathcal{F}$ of NAE-SAT, we build a graph $G(\mathcal{F})$ as follows.

To the clauses we associate vertex disjoint triangles; each vertex corresponds to one of the literals of the clause. For each variable x, vertex disjoint paths P_x are added to the graph as follows. Let $C_1, \dots, C_k$ be the clauses in which x or its negation occur, the path P_x is defined with vertices $v_{x_1} \dots v_{x_{2k}}$ (in this order). The path P_x and the triangles are joined with the following rule: if C_i contains x (resp., $\overline{x}$), we add the edge from the vertex of the triangle representing x to $v_{x_{2i-1}}$ (resp., to $v_{x_{2i}}$). This construction is clearly polynomial in the size of $\mathcal{F}$, and it is easy to verify that $G(\mathcal{F})$ is 2-clique-colorable if and only if $\mathcal{F}$ is not-all-equal satisfiable. Furthermore, $G(\mathcal{F})$ is of maximum degree 3. □

Because of the nature of the clique-coloring problem, the NP-completeness of the 2-clique-coloring problem does not immediately imply the NP-completeness of the k-clique-coloring problem (for any fixed $k \geq 2$). Nevertheless it is true; here is a simple reduction.

COROLLARY 1. *For any fixed $k \geq 2$, the k-clique-coloring problem is NP-complete.*

Proof. Let G be an instance of the k-clique-coloring problem. Add a copy of the $(k+2)$-chromatic Mycielski graph G_{k+2}. Remove an edge incident to x_{k+2} (we use the notation given in the introduction), and replace x_{k+2} by $|V(G)|$ copies of x_{k+2}. Pairing these copies of x_{k+2} with the vertices of G, we obtain a new graph G'. Observe now that in any $(k+1)$-coloration of G', all copies of x_{k+2} have the same color. Hence a $(k+1)$-clique-coloration of G' yields a k-clique-coloration of G, which completes the reduction. □

3. How to clique-color a graph? It is not difficult to provide clique-coloration of a graph: just color every vertex with a different color; a coloration of the graph is also a proper clique-coloration, etc. However, the clique-chromatic number is typically much smaller than the chromatic number. For instance, for perfect graphs the chromatic number is ω and the clique-chromatic number is conjectured to be a constant, maybe 3!

We need heuristics that may provide better estimates than the chromatic number. Besides the difficulty of coloring with a small number of colors, it is also difficult to realize that a procedure is good, since by Theorem 1 we cannot even check whether a partition of the vertices is a clique-coloration.

However, certain constructions inherently guarantee that the result is a proper coloration, and at the same time the number of occurring colors can be bounded in a helpful way. We present in this section three such frameworks. These are meant to be used more as frameworks than algorithms: in the realizations queues can be broken in various ways, and this arising freedom will be exploited in the particular procedures we will present later.

A neighborhood-coloration is any clique-coloration obtained by the following greedy framework.

Neighborhood coloring.

INPUT: Graph $G = (V, E)$ and $\mathcal{H} \subseteq \mathcal{H}(G)$.

0. In each iteration, the algorithm updates the set D of "considered" vertices and the set L of "colored" vertices, $D \subseteq L$. Initially set $D := \emptyset$, $L := \emptyset$.

While not all the vertices are colored do the following:

1. Choose $v \in V \setminus D$, and consider v.

2. If $v \notin L$, then assign to v a color which does not occur in $N(v)$; $L := L \cup \{v\}$.

3. Let c be a color different from all colors occurring among the neighbors of vertices in $N(v) \setminus L$. Assign to all vertices in $N(v) \setminus L$ the color c.

4. Update: $D := D \cup \{v\}$, $L := L \cup N(v)$.

Lemma 1. *The coloration found by the algorithm is a clique-coloration of G.*

Remark. At each iteration the set of considered vertices dominates the set of colored vertices, so that the set D obtained at the end of the algorithm is a dominating set of G; that is, $N[D] = V$.

The order in which the vertices are considered, or the free choices for the colors, for instance, for color c, will be replaced by particular rules in more specific coloring procedures.

The next lemma shows that if a graph admits a certain partition of the vertices, then it is k-clique-colorable. A clique-coloration obtained by the way described in the proof of Lemma 2 will be called a *partition coloration.*

Lemma 2. *Let $G = (V, E)$ be a graph and $k \in \mathbb{N}$, $k \geq 2$.*

If G admits a partition $\{V_1, \dots, V_p\}$ of V such that

- *$G(V_i)$ is k-clique-colorable, and V_i has a border-guard in G ($i = 1, \dots, r \leq p$);*
- *$G(V_i)$ ($i = r+1, \dots, p$) does not contain a maximal clique of G;*
- *the graph H obtained by identifying the vertices of each V_i (denote the new vertices by x_i, $i = 1, \dots, p$) has $\chi(H) \leq k$;*

then G is k-clique-colorable.

Proof of Lemma 2. Consider a k-coloration $c_H : V(H) = \{x_1, \dots, x_p\} \longrightarrow \{1, \dots, k\}$ of H and also a k-clique-coloration $c_i : V_i \longrightarrow \{1, \dots, k\}$ of $G(V_i)$ ($i = 1, \dots, r$).

By assumption V_i has a border-guard v_i in G ($i = 1, \dots, r$). We can suppose that $c_i(v_i) = c_H(x_i)$ (otherwise we interchange two colors in the coloration of $G(V_i)$). Furthermore, for $i = r+1, \dots, p$ we define $c_i(v) = c_H(x_i)$ for all $v \in V_i$. Define for $v \in V(G)$ $c(v) := c_i(v)$ if $v \in V_i$.

Now let Q be a maximal clique of G. If Q is contained in some V_i, then by the assumption $i \le r$ and $c(q) = c_i(q)$ for all $q \in Q$. Therefore, at least two colors occur in Q. If Q is not contained in some V_i, then say $Q \cap V_i \neq \emptyset \neq Q \cap V_j$. Let $v_i \in Q \cap V_i$ (resp., $v_j \in Q \cap V_j$) be an arbitrary vertex in $Q \cap V_i$ (resp., $Q \cap V_j$) for $i \ge r$ (resp., $j \ge r$).

Clearly, $v_i, v_j \in Q$. Since $c(v_i) = c_H(x_i) \neq c_H(x_j) = c(v_j)$ because of $x_i x_j \in E(H)$, two different colors do occur in Q. □

A third simple but useful method is presented in the following lemma. A pair (d, D) is called a *dominating pair* if $d \in V$, $D \subseteq N(d)$, and any maximal clique K of G containing d satisfies $K \cap D \neq \emptyset$. The following lemma shows that such a pair can be useful for our coloring problem.

LEMMA 3 (dominating pair lemma). *Let (d, D) be a dominating pair, and let k be a nonnegative integer with $|D| < k$. If $\mathcal{H}(G-d)$ is k-colorable, then so is $\mathcal{H}(G)$.*

Proof. Let c be a k-coloration of $\mathcal{H}(G-d)$. Since $k > |D|$, there exists a color i that does not occur in D. Let $c' : V \to \{1, 2, \dots, k\}$, with $c'(v) = c(v)$ for all $v \in G-d$ and $c'(d) = i$. Since c is a k-coloration of $\mathcal{H}(G-d)$, it is sufficient to check that any maximal clique K which contains d is not monocolored by c'. By definition of a dominating pair, there exists a vertex $v \in K \cap D$. By the choice of i, we have $c'(d) = i \neq c(v) = c'(v)$. Thus c' is a k-coloration of $\mathcal{H}(G)$. □

Let G be a graph with the property that every induced subgraph contains a vertex u whose neighborhood has at most k connected components, each of which is a clique. A direct consequence of the dominating pair lemma is that G is $k+1$-clique-colorable.

4. Rough general bounds.

In this section we estimate the clique-chromatic number with some other graph parameters.

Recall that a *dominating set* D is a subset of V such that $N[D] = V$. The *domination number* $\gamma(G)$ of a graph G is the smallest cardinality of such a set. Note that $\gamma(G)$ is always smaller than or equal to the stability number $\alpha(G)$.

We assume G to be connected, leaving to the reader the trivial extension of the following theorem to graphs with several connected components.

THEOREM 3. *If $G = (V, E)$ is a connected graph, then $\kappa(G) \le \gamma(G) + 1$, and if $\kappa(G) = \gamma(G) + 1$, then every dominating set D of minimum size is a stable set, and one of the following holds:*

- *$|D| < \alpha(G)$,*
- *D is a set of two nonadjacent vertices of $G = C_5$,*
- *$|D| = 1$ and $G = K_n$, $n \ge 2$.*

Proof of Theorem 3. Let $D = \{x_1, \dots, x_k\}$ be a dominating set of G, and $n := |V(G)|$. If there exists $a, b \in D$, $ab \in E(G)$, suppose $x_k = b$. Apply a neighborhood coloring with the following specifications: the order of considering the vertices is $x_1, \dots, x_k$; in the ith iteration ($i = 1, \dots, k$), if x_i is not yet colored, color it with color 1; moreover, for $i = 1, \dots, k-1$, color the not yet colored vertices of $N(x_i)$ with color $i+1$; if $c(x_k) \neq 1$, then color $N(x_k) \setminus \cup_{j=1}^{k-1} N[x_j]$ with color 1, otherwise with color $k+1$. It can be checked immediately that the defined colors are allowed, and the number of colors is $k+1$ only if D is a stable set. More exactly, we have the following claims.

Claim 1. If $\kappa(G) = k+1$, then D is a maximal stable set of minimum size.

Indeed, if there exists a maximal stable set D' of smaller size $k' := |D'| < k$, then it is also a dominating set. Hence $\kappa(G) \le k' + 1 \le k$, as required.

Assume now that $k = \alpha(G)$.

Claim 2. If $\kappa(G) = k + 1$, $k = \alpha(G) \le 2$, then either $G = C_5$, or $G = K_n$, $n \ge 2$.

Indeed, if $k = \alpha = 1$, then $G = K_n$. Let now $k = \alpha = 2$. We prove by induction on the number of vertices that $\kappa(G) = 2$, unless $G = C_5$.

Let a and b be two nonadjacent vertices; then because of $\alpha < 3$, $N[a] \cup N[b] = V(G)$.

If we can 2-clique-color the subgraph N_{ab} induced by $N(a) \cap N(b)$, then we extend this coloration to all G: define $c(v) := 1$ if $v \in \{a\} \cup N(b) \setminus N(a)$, and $c(v) := 2$ if $v \in \{b\} \cup N(a) \setminus N(b)$. If Q is a maximal clique of G and, say, $c(q) = 1$ for all $q \in Q$, then all vertices of $Q \setminus a$ are adjacent to b. Since $c(b) = 2$ it follows that $a \in Q$. But then $Q \setminus a$ is a maximal clique of N_{ab}, and, since c is a 2-clique-coloration of N_{ab}, $Q \setminus a$ is a single vertex, v. If $\{b, v\}$ is not a maximal clique, then by giving color 2 to v we get a 2-clique-coloration of G. Else v is adjacent only to a and b, and so, since $\alpha = 2$, $V(G) \setminus \{a, b, v\}$ is a clique. We may assume that $N_{ab} \setminus \{v\}$ is empty and that $N(a) \setminus N(b)$ and $N(b) \setminus N(a)$ are nonempty, since, else, there exists a dominating edge in G and hence, by Claim 1, a 2-clique-coloration of G. In case a or b has at least two neighbors distinct from v, then let w be one of those, give color 1 to a, b, and w, and give color 2 to all the other vertices: this a 2-clique-coloration of G. The only remaining case is when $|N(a) \setminus N(b)| = |N(b) \setminus N(a)| = 1$; then $G = C_5$.

We now assume that N_{ab} has no 2-clique-coloration. Thus by induction hypothesis, at least one connected component of N_{ab} induces a C_5. Since $\alpha = 2$, we have $N_{ab} = C_5$. Label $v_1, \dots, v_5$ its vertices in the cyclic order. If $N(a) \setminus N(b) = N(b) \setminus N(a) = \emptyset$, then G is 2-clique-colorable; else fix a vertex v in, say, $N(a) \setminus N(b)$. Since $\alpha(G) = 2$, v is adjacent either to v_1 or to v_3, say v_1, and v is adjacent either to v_2 or to v_5, say v_2. Now give color 1 to a, v_1, v_2, v_4, and all the vertices in $N(b) \setminus N(a)$, and give color 2 to all the other vertices: this a 2-clique-coloration of G.

The claim is now proved.

To finish the proof of Theorem 3, suppose that $k \ge 3$ and that D is a stable set of cardinality $k = \alpha(G)$. In the above constructed neighborhood coloring, let x_{k-2}, x_{k-1}, x_k be the three pairwise nonadjacent vertices colored last. The neighborhood coloring assigns colors $c(x_{k-2}) = c(x_{k-1}) = c(x_k) = 1$ and new colors $k - 1$, k, $k + 1$ to the set of their not-yet-colored neighbors.

Claim 3. The graph induced by vertices of color $k - 1$, k, $k + 1$ and x_{k-2}, x_{k-1}, x_k can be 3-clique-colored.

The claim finishes the proof of the theorem. Indeed, choose the three colors to be 1, $k - 1$, and k to get a k-clique-coloration of G. (The colors $k - 1$ and k do not occur previously, and all previously colored vertices of color 1 are nonadjacent to the vertices that are present in the claim.)

To prove Claim 3, we can suppose $k = 3$; then the notation is simplified, and we only have to prove $\kappa(G) \le 3$.

If $G - N[v]$ is not a C_5 for some $v \in V(G)$, then by Claim 2 it can be colored with colors 1 and 2; completing this coloration with $c(v) := 1$ and $c(x) := 3$ if $x \in N(v)$, the statement is proved.

Suppose now that $G - N[v]$ is a C_5 for all $v \in V(G)$. Then G is $n - 6$-regular. If there is no triangle in G, then $N(v)$ is a stable set for all $v \in V(G)$, and therefore $n - 6 \le 3$. The equality holds here, because if G is 2-regular, then $G - N[v]$ cannot be a C_5 for all $v \in V(G)$. But if the equality holds, then the number of edges with exactly one endpoint in $N[v]$ is, on one hand, $2|N(v)| = 6$ and, on the other hand, 5 (because there is exactly one such edge for every vertex of $G - N[v]$).

So G has a triangle. Let $ab \in E(G)$ be one of its edges. If $\{a, b\}$ is a dominating set, then we can 2-clique-color G by Claim 1. Let us suppose that v is adjacent neither to a nor to b. Since $G - N[v]$ is a C_5 containing the edge ab, where ab is contained in a triangle of G, the following coloration is correct: $c(v) := c(a) := c(b) := 1$, $c(x) := 2$ if $x \in N(v)$, and the remaining three vertices forming a path in the C_5 can be colored $3, 1, 3$. □

Remark that for any integer k, a path P_{3k} on $3k$ vertices has a dominating number equal to k and $\kappa(P_{3k}) = 2$.

On the other hand, Mycielski's graphs provide an infinite class of triangle-free graphs G_k for which $\kappa(G_k) = \chi(G_k) = \gamma(G_k) + 1 = k$ (for $k \geq 4$ the first case of the theorem holds, for $k = 3$ the second, and for $k = 2$ the third). Let $D_2 = \{v\}$, where v is either vertex of G_2, and define $D_k := D_{k-1} \cup \{x_k\}$ (we use the notation given in the introduction). By construction, D_k is a dominating set of G_k and $|D_k| = k - 1$. By the theorem, and since $\kappa(G_k) = \chi(G_k) = k$, we have that $\gamma(G_k) = k - 1$, and it follows that D_k is a maximal stable set of minimum size (and not maximum as soon as $k \geq 4$).

Corollary 2. *For any graph $G \neq C_5$ with $\alpha(G) \geq 2$, we have $\kappa(G) \leq \alpha(G)$.* □

This first corollary sharpens Theorem 2 in [7]. Indeed, it is stated there that $\kappa(G) \leq \alpha(G) + 1$ and the strict inequality holds for C_5-free noncomplete graphs.

Corollary 3. *For any graph G of order n, we have $\kappa(G) \leq 2\lceil\sqrt{n}\,\rceil$.*

Proof. Let $D = \{v_1, \dots, v_k\}$ be a subset of k vertices with the following properties:

- $|N(v_1)| \geq \sqrt{n}$,
- $|N(v_i) - (\cup_{j<i} N[v_j])| \geq \sqrt{n}$ for $i = 2, \dots, k$,
- any vertex $v \in V(G)$ satisfies $|N(v) - N[D]| < \sqrt{n}$.

Note that D can be empty. Since D is a dominating set of $N[D]$, and $|D| < \sqrt{n}$, by Theorem 3, we can clique-color the subgraph induced by $N[D]$ with $\lceil\sqrt{n}\,\rceil$ colors, say $\{1, \dots, \lceil\sqrt{n}\,\rceil\}$.

On the other hand, in the subgraph induced by $V \setminus N[D]$ the degree of every vertex is strictly smaller than $\sqrt{n}$, so we can color this subgraph with $\lceil\sqrt{n}\,\rceil$ colors, say $\{\lceil\sqrt{n}\,\rceil + 1, \dots, 2\lceil\sqrt{n}\,\rceil\}$, by a sequential algorithm. This coloration is a clique-coloration too. □

This bound is not best possible: Kotlov [9] proved that $\kappa(n) \leq \lfloor\sqrt{2n}\rfloor$. We do not even know whether the maximum of the clique-chromatic number for graphs on n vertices divided by $\sqrt{2n}$ is a constant or tends to 0.

Theorem 4. *Let $G = (V, E)$ be a graph and q be an integer, $q > 1$. Then the hypergraph $\mathcal{H}_q := \{K \in \mathcal{H}(G) : |K| \geq q\}$ is $\lceil\frac{\chi(G)}{q-1}\rceil$-colorable.*

Proof. Let $k := \lceil\chi(G)/(q-1)\rceil$. Let $S_1, \dots, S_{\chi(G)}$ be the color classes of a $\chi(G)$-coloration of G. For $i = 1, \dots, k$, we consider the union of $q - 1$ color-classes: $C_i = \bigcup_{j=(i-1)(q-1)+1}^{i(q-1)} S_j$ if $i = 1, \dots, k - 1$, and $C_k = \bigcup_{j=(k-1)(q-1)+1}^{\chi(G)} S_j$.

Observe that $\omega(C_i) < q$ for every $i = 1, \dots, k$. Thus, the coloration c, defined by $c(x) = i$ if $x \in C_i$, is a k-coloration of $\mathcal{H}_q$. □

Corollary 4. *If G is an arbitrary graph, then $(\kappa - 1)(\bar{\kappa} - 1) \leq 2\min\{\chi, \bar{\chi}\} - 2$.*

Proof. Let k be the size of a smallest maximal stable set of G. Since a maximal stable set of G is a dominating set of G, by Theorem 3, we have that $\kappa(G) - 2 \leq k - 1$.

By the choice of k, we have that any maximal clique of $\overline{G}$ has size at least k. If $k > 1$, by Theorem 4, we obtain $\kappa(\overline{G}) - 1 \leq \frac{\chi(\overline{G})-1}{k-1}$. Multiplying the two inequalities, we obtain $(\kappa - 2)(\bar{\kappa} - 1) \leq \bar{\chi} - 1$.

If $k = 1$, then $\kappa = 2$ and trivially $(\kappa - 2)(\bar{\kappa} - 1) \leq \bar{\chi} - 1$.

In both cases we get $(\kappa - 1)(\bar{\kappa} - 1) \leq \bar{\chi} + \bar{\kappa} - 2 \leq 2(\bar{\chi} - 1)$.

Applying this again after interchanging the role of G and $\bar{G}$, we get the claim. □

This bound can be sharpened under various assumptions. For instance, if κ or $\bar{\kappa}$ are close to χ or $\bar{\chi}$, like for Mycielski graphs (see section 1), if $\kappa = \chi$, then $\bar{\kappa} \leq 3$. (In fact, for Mycielski graphs the statement "$\bar{\kappa} = 2$ except for $G_3 = C_5$" is easy to prove directly.) The bound can also be refined using other parameters: as Kotlov [9] noticed, $(\kappa - 1)(\bar{\kappa} - 1) \leq \frac{k}{k-1}(\bar{\chi} - 1)$ if $k > 1$.

5. Claw-free and perfect graphs. In this section we study $\kappa(G)$ and $\kappa(\bar{G})$ when G is a claw-free or a perfect graph or both.

If G is a perfect graph, then we have $\kappa(G) \leq \chi(G) = \omega(G)$. Applying also Corollary 2, if G is not a complete graph, then we have $\kappa(G) \leq \min\{\alpha(G), \omega(G)\}$. (This is better than the bound of Corollary 4 only if $\bar{\kappa} = 2$.) Moreover, when G is perfect, $\alpha(G)$ and $\omega(G)$ can be computed in polynomial time [6].

Furthermore, it seems that in perfect graphs not only the maximum cliques but also the maximal cliques behave well from the viewpoint of clique-colorations. A consequence could be that there exists a constant C such that $\mathcal{H}(G)$ is C-colorable for a perfect graph G; that is, Question 1 has a positive answer. We prove that such a C exists for some classes of perfect graphs.

For example, the hypergraph of maximal cliques of a strongly perfect graph G (defined by the property that every induced subgraph of G contains a stable set intersecting all maximal cliques) is obviously 2-colorable: indeed, color a stable set intersecting all maximal cliques of G with one color and the rest of the vertices with another color.

Note that $\kappa(G)$ can be greater than 2, even for a perfect graph G (see Figure 5.1).

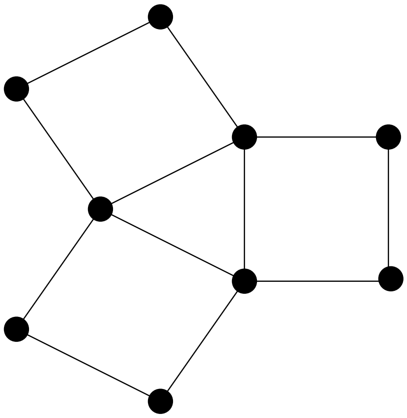


FIG. 5.1. *The clique-hypergraph of this perfect graph is clearly not 2-colorable since it contains edges of C_9 as hyperedges.*

We saw in the introduction that the clique-chromatic number of claw-free graphs or even of line-graphs is not bounded. The following theorem shows that triangles are the only source of difficulty.

We do not know the complexity of clique-coloring line-graphs of graphs optimally. Observe that in the case of line-graphs, it is easy to check whether a given coloration is correct since all maximal cliques of a line-graph $L(G)$ are either stars or triangles of G, and therefore the number of maximal cliques is small (bounded by a polynomial of the number of vertices).

A multigraph is a graph that may contain an arbitrary number of parallel edges.

THEOREM 5. *Let G be a multigraph, $\mathcal{H} = (V, \mathcal{E})$ where $V := E(G)$, and $\mathcal{E}$ is the collection of stars of G. Then $\chi(\mathcal{H}) \leq 3$. Moreover, $\chi(\mathcal{H}) = 3$ if and only if G has a component which is an odd circuit.*

Proof. Without loss of generality, assume that G is connected. Let G' be obtained from G by adding to it a perfect matching M of its odd-degree vertices, if any. Let T be an Eulerian tour of G'. Color the edges of T alternatively black and white, starting at a vertex of degree at least four (if any) or with an edge of M (if any). If there is such a vertex or such an edge, then this coloring induces a proper 2-coloration of $\mathcal{H}$. Else, G is a cycle, and this 2-coloration of $\mathcal{H}$ is not proper if and only if G is an odd cycle. □

We are highly indebted to Kotlov [9] for short-cutting most of our original proof.

For complements of claw-free graphs, the following simple bound holds.

THEOREM 6. *Let $2 \leq k \leq \alpha(G)$. If G is $K_{1,k}$-free, then $\kappa(\bar{G}) \leq k$.*

Proof. Since $k \leq \alpha(G)$, there exists a stable set $S \subseteq V(G)$, $|S| = k$. Since G is $K_{1,k}$-free, S induces a dominating clique (not necessarily maximal) of $\overline{G}$. We achieve the proof of Theorem 6 by applying Theorem 3. □

Notice that the complements of Mycielski's graphs are $K_{1,3}$-free, showing that the condition $k \leq \alpha(G)$ in the preceding theorem is necessary.

We have now arrived at the most difficult result of this paper: we determine the clique-chromatic number of claw-free perfect graphs.

THEOREM 7. *If G is a claw-free perfect graph, then $\mathcal{H}(G)$ is 2-colorable.*

By Theorem 6 any graph which is the complement of a claw-free graph of stability number at least 3 is 3-clique-colorable even if it is not perfect. On the other hand, we saw that line-graphs (which are, of course, claw-free) may have arbitrary large clique-chromatic number, unless they arise from triangle-free graphs.

In [7] it is proved that the hypergraph of *maximum* cliques of a claw-free graph is 2-colorable if and only if it does not contain an odd hole. A common feature of the proof of [7] and our proof below is the use of Ben Rebea's lemma (as cited in [2]); however, an essential difference is that the main part of our proof is the perfect case.

COROLLARY 5. *If G is a claw-free graph without an odd hole, then $\kappa(G) \leq 2$.*

Proof of Corollary 5. Let G be claw-free without an odd hole. If $\alpha(G) \leq 2$, then by Corollary 2, $\kappa(G) \leq 2$.

If $\alpha(G) \geq 3$, then G is perfect because of the following: by Ben Rebea in [2] a connected claw-free graph G with $\alpha(G) \geq 3$ containing an odd antihole also contains an odd hole; Parthasaraty and Ravindra [18] proved that a claw-free graph with neither an odd hole nor an odd antihole is perfect.

Since G is perfect, Theorem 7 can now be applied. □

In order to prove Theorem 7, we use the structural property of claw-free graphs explored by Chvátal and Sbihi [2] and Maffray and Reed [14].

Chvátal and Sbihi [2] defined two special classes of claw-free perfect graphs: the *elementary graphs* and *peculiar graphs.* A graph is called elementary if its edges can be colored with two colors such that every induced P_3 (chordless path on three vertices) has its two edges colored differently. Clearly elementary graphs are claw-free, but not vice versa, as C_5 shows. A graph is called peculiar if it can be obtained as follows: take three pairwise vertex-disjoint cobipartite graphs; call them $(A_1, B_2), (A_2, B_3), (A_3, B_1)$, such that each of them has at least one pair of non-adjacent vertices; add all edges between every two of these cobipartite graphs; then add three cliques Q_1, Q_2, Q_3 that are pairwise disjoint and disjoint from the A_i's and B_i's; add all the edges between Q_i and $A_j \cup B_j$ for $j \neq i$; there is no other edge in the graph. Chvátal and Sbihi [2] proved that every claw-free perfect graph can be decomposed via clique-cutsets into indecomposable graphs that are either peculiar or elementary.

THEOREM 8 (see [2]). *If G is a claw-free perfect graph without a clique cutset, then G is either elementary or peculiar.*

The structure of elementary graphs was determined by Maffray and Reed in [14] as follows. An edge is called *flat* if it does not lie in a triangle. Let xy be a flat edge of a graph G and $(X, Y; F)$ be a cobipartite graph disjoint from G and containing at least one edge with one extremity in X and the other in Y. We obtain a new graph from $G - \{x, y\}$ and $(X, Y; F)$ by making the union of their sets of vertices and edges and adding all possible edges between X and $N_G(x) \setminus \{y\}$ and between Y and $N_G(y) \setminus \{x\}$. This is called *augmenting* the flat edge xy with the cobipartite graph $(X, Y; F)$. The result of augmenting a set of pairwise independent (nonincident) flat edges $e_1, \ldots, e_h$ successively is called an *augmentation* of G.

THEOREM 9 (see [14]). *A graph G is elementary if and only if it is an augmentation of the line-graph of a bipartite multigraph.*

Proof of Theorem 7. We now prove Theorem 7 through several lemmas.

LEMMA 4. *If G is an elementary graph, then $\mathcal{H}(G)$ is 2-colorable.*

Proof of Lemma 4. For line-graphs of bipartite multigraphs the statement follows from Theorem 5. Furthermore, if G has a 2-clique coloration, the graph obtained by augmenting a flat edge xy with $B = (X, Y; F)$ still has a 2-clique-coloration: keep the same color for all vertices of $G - \{x, y\}$; choose an edge ab of B with $a \in X$ and $b \in Y$; and give color 1 to a and to all vertices in $Y \setminus \{b\}$ and color 2 to b and to all vertices in $X \setminus \{a\}$. □

Using previous results, it is also not difficult to check the following.

LEMMA 5. *If G is a peculiar graph, then $\mathcal{H}(G)$ is 2-colorable.*

Proof of Lemma 5. Let $G = (V, E)$ be a peculiar graph composed of (A_1, B_2), (A_2, B_3), (A_3, B_1), Q_1, Q_2, Q_3 as in the definition of a peculiar graph. Let $a \in A_1$ and let $b \in B_3$ (by definition all the A_i's, B_i's are nonempty). It is easy to verify that the edge ab is dominant, and hence by Theorem 3 we obtain that $\mathcal{H}(G)$ is 2-colorable. □

LEMMA 6. *If G is a claw-free graph and Q is a clique which is a minimal cutset, then $G - Q$ has two components; denote their set of vertices V_1 and V_2, and at least one of the following holds:*

(a) *Either for $i = 1$ or for $i = 2$ both V_i and $V \setminus V_i$ have a border-guard.*

(b) *Both V_1 and V_2 have a border-guard.*

(c) *Both $V_1 \cup Q$ and $V_2 \cup Q$ have two border-guards.*

Proof of Lemma 6. Since Q is a minimal cutset, every $q \in Q$ has a neighbor in all the components. Since G is claw-free, $G - Q$ has two components, and $N(q) \cap V_i$ is a clique for all $i = 1, 2$ and all $q \in Q$.

Claim 1. For all $a, b \in Q$, either $N(a) \cap V_1 \subseteq N(b) \cap V_1$ or $N(a) \cap V_2 \subseteq N(b) \cap V_2$.

Indeed, if not, let $a_i \in N(a) \cap V_i \setminus (N(b) \cap V_i)$ $(i = 1, 2)$. Clearly, a, b, a_1, a_2 induce a claw, a contradiction.

Claim 2. Either there exists a border-guard in V_1, or there exist two distinct border-guards in $V_1 \cup Q$.

Indeed, suppose the first possibility does not hold. Then there are $a \neq b \in Q$ so that $N[a] \cap V_1$ and $N[b] \cap V_1$ are not equal, and they are both inclusionwise minimal among $N[q] \cap V_1$ ($q \in Q$). (If there were a unique inclusionwise minimal $N[q] \cap V_1$ ($q \in Q$), then any $v_1 \in N[q] \cap V_1$ would be a border-guard of V_1.)

Since neither $N[a] \cap V_1$ nor $N[b] \cap V_1$ contains the other, by Claim 1 both $N[a] \cap V_2 \subseteq N[b] \cap V_2$ and $N[b] \cap V_2 \subseteq N[a] \cap V_2$ hold; that is, $N[a] \cap V_2 = N[b] \cap V_2 =: N_2$.

Now by the minimal choice of $N[a] \cap V_1$ and of $N[b] \cap V_1$, $N[q] \cap V_1$ for any $q \in Q$ cannot be a subset of both. So by Claim 1, $N[q] \cap V_2 \subseteq N_2$ for all $q \in Q$. Since $B(V_1 \cup Q) = Q \cup N_2$, we proved that both a and b are border-guards of $V_1 \cup Q$ and the claim is proved.

To finish the proof of Lemma 6, note that by symmetry, Claim 2 also holds if we replace 1 by 2. From these two variants of Claim 2 we get that one of the following cases holds:

– Both V_1 and V_2 have a border-guard, and then each of these is adjacent with every vertex in Q. So Q is not a maximal clique, and "b" of the lemma holds.

– Both $V_1 \cup Q$ and $V_2 \cup Q$ have two border-guards, and then we have "c."

– V_1 and $V_2 \cup Q$ have border-guards or V_2 and $V_1 \cup Q$ have border-guards. This is just "a." □

The proof of Theorem 7 works by induction on $|V|$. Let $G = (V, E)$ be a claw-free perfect graph. If G has one, two, or three vertices, then clearly $\mathcal{H}(G)$ is 2-colorable. Suppose now that G has n vertices and that the theorem has been proved for any claw-free perfect graph with less than n vertices. If G is either elementary or peculiar, then, by Lemmas 4 and 5, $\mathcal{H}(G)$ is 2-colorable. So by Theorem 8, we may assume that G has a clique cutset.

We can now finish the proof of Theorem 7 by applying the idea of Lemma 2 in a very simple special case.

If Lemma 6(a) holds for say $i = 1$, by the induction hypothesis, we can 2-clique-color $G(V_1)$ and $G(V \setminus V_1)$. Without loss of generality, we may assume that the border-guard of V_1 has a different color from that of $V \setminus V_1$. Every maximal clique of G is contained either in V_1 or in $V \setminus V_1$, or contains both border-guards. In any case, both colors occur in it.

If Lemma 6(b) holds, then by the induction hypothesis, we can 2-clique-color $G(V_1)$ and $G(V_2)$. Without loss of generality, we may assume that both their border-guards have color 1. Color all vertices of Q with color 2. Since every maximal clique of G is contained in V_1 or V_2 or contains a border-guard and a vertex of Q, we defined a 2-clique-coloration.

Finally, if Lemma 6(c) holds, then color Q so that the two border-guards of $V_1 \cup Q$, and also those of $V_2 \cup Q$, have different colors, and otherwise arbitrarily. We complete this coloration by a 2-clique-coloration of $G(V_1)$ and $G(V_2)$. Now every maximal clique of G is contained in V_1 or in V_2, or for some $i \in \{1, 2\}$ it contains both border-guards of $V_i \cup Q$. □.

Note that the proof of Theorem 5 is algorithmic; moreover, either it reduces the clique-coloration of G into the clique-coloration of two smaller graphs or the graph itself is easy to color.

Using the following ingredients, the proof provides a way of 2-clique-coloring an arbitrary claw-free perfect graph G in polynomial time:

– Whitesides's algorithm [23] that finds a clique cutset;

– Chvátal and Sbihi's Theorem 8 [2];
– Maffray and Reed's canonical decomposition algorithm of an elementary graph into a line-graph of a bipartite graph and some augmentations [14];
– checking for border-guards is polynomial (obvious);
– the number of graphs occurring through the decomposition can be bounded by a polynomial of the number of vertices of the input graph. (These graphs are not the same as in Chvátal and Sbihi's algorithm for recognizing claw-free perfect graphs, since the clique-cutset is not left in both two decomposing graphs.)

Furthermore, this algorithm uses only the graph G and not a list of its maximal cliques.

Diamond-free perfect graphs constitute another interesting class of perfect graphs (a diamond is a K_4 minus an edge). It is known [22, 19] that a diamond-free graph is perfect if and only if it does not contain an odd hole. Unfortunately we cannot prove $\kappa \leq 3$ for this class. This is somewhat frustrating, because Tucker [22] proved that a diamond-free perfect graph has a vertex which is contained in at most two maximal cliques of size at least 3, which implies the following.

Proposition 1. *The hypergraph of maximal cliques of size at least* 3 *of a diamond-free perfect graph is* 3*-colorable. In particular, if G is a diamond-free perfect graph without flat edges, then $\kappa(G) \leq 3$.*

The conjecture $\kappa \leq 3$ for diamond-free perfect graphs (equivalently diamond- and odd-hole-free graphs) could contain many of the difficulties of coping with odd-hole-free graphs in general. We wonder whether the clique-chromatic number of odd-hole-free graphs could be bounded as well: we also do not know of any odd-hole-free graph with clique-chromatic number greater than three.

6. Generalized split graphs. A graph G is a *generalized split graph* if either G or the complement of G has a vertex partitioned into sets A, B_i $(1 \leq i \leq k)$ so that A and all B_i's span complete graphs and there are no edges between B_i and B_j if $i \neq j$. Generalized split graphs are perfect and have been introduced in the paper of Prömel and Steger [20]; this class plays a crucial role in their proof of the asymptotic version of the strong perfect graph conjecture: almost all Berge graphs are perfect. In fact, they proved in [20] that almost all C_5-free graphs are generalized split graphs. ("Almost all" means here that the ratio of the number of labelled n-vertex C_5-free graphs to the number of n-vertex generalized split graphs tends to one if n tends to infinity.) Therefore any property of generalized split graphs holds for almost all perfect graphs. In our case the property in question is the chromatic number of the clique hypergraph.

Theorem 10. *The clique-hypergraph of a generalized split graph is* 3*-colorable.*

Proof. Assume that G is a generalized split graph. If the complement of G has the required partition into A, B_i's, then a proper coloration for the maximal cliques of G is trivial: the vertices of A are colored with color 1, the vertices of B_1 are colored with color 2, and the vertices in all other B_i's (if there are any) are colored with color 3.

If G has the required partition, then two cases are considered. If $|A| \leq 1$, then we color the B_i's with colors 1 and 2 so that each of them with at least two vertices gets both color 1 and color 2, and if A is nonempty, we color it with color 3. Finally, if $|A| > 1$, a fixed vertex $x \in A$ is colored by color 2, all other vertices of A are colored with color 3, the sets B_i with one vertex are colored with color 1, and any set B_i with at least two vertices is colored using the same rule: if x is adjacent to all vertices of

B_i, then color all vertices of B_i with color 1; otherwise, a fixed vertex of B_i which is not adjacent to x is colored with color 2 and all other vertices of B_i are colored with color 1. It is straightforward to check that under this coloration every maximal clique of G gets at least two colors. □

It is worth noting that the theorem is sharp in the sense that there are generalized split graphs with 3-chromatic clique-hypergraphs, for instance, the graph in Figure 5.1.

The result of Prömel and Steger [20] mentioned above yields the following corollary, which is an asymptotic answer to Question 1.

Corollary 6. *Almost all perfect graphs are* 3*-clique-colorable.*

7. Open problems. In Theorem 1, we proved that Maximal clique containment is NP-complete for the complements of $K_{1,4}$-free graphs. It is therefore natural to first ask the following question.

Question 2. Is Maximal clique containment polynomially solvable for the complements of $K_{1,3}$-free graphs?

Since it is NP-complete to compute the chromatic number of a triangle-free graph [12], it is NP-complete to compute the clique-chromatic number of a complement of a $K_{1,3}$-free graph. Nevertheless, we know by Theorem 6 that $\chi(\mathcal{H}(\bar{G})) \leq 3$ when G is $K_{1,3}$-free and $\alpha(G) \geq 3$. Hence we should ask the next question.

Question 3. Is it NP-complete to determine whether $\bar{G}$ is 2-clique colorable when G is $K_{1,3}$-free?

We saw that it is NP-complete to determine whether a graph of maximum degree 3 is 2-clique-colorable. Moreover, Corollary 5 gives that any $K_{1,3}$-free graph with no odd hole is 2-clique colorable.

Question 4. Is it NP-complete to determine whether G is 2-clique colorable when G is $K_{1,3}$-free?

Most of our results concern classes of graphs defined by forbidden configurations. Thus it would be interesting to study hereditary properties of the clique-chromatic number of a graph. Hoàng and McDiarmid in [7] studied such questions. Concerning the complexity aspect, we ask the following.

Question 5. What is the complexity of deciding whether a graph and all its induced subgraphs can be 2-clique-colored?

Acknowledgments. We are indebted to Andrei Kotlov for a thorough reading of the manuscript, many precious remarks, and detections of flaws, and to Colin McDiarmid for a very interesting discussion opening to us new aspects of clique colorations. We are also very thankful to an anonymous referee for several helpful comments.

REFERENCES

[1] Y. Caro, A. Sebő, and M. Tarsi, *Recognizing greedy structures*, J. Algorithms, 20 (1996), pp. 137–156.

[2] V. Chvátal and N. Sbihi, *Recognizing claw-free Berge graphs*, J. Combin. Theory Ser. B, 44 (1988), pp. 154–176.

[3] D. Duffus, H. A. Kierstead, and W. T. Trotter, *Fibres and ordered set coloring*, J. Combin. Theory Ser. A, 58 (1991), pp. 158–164.

[4] D. Duffus, B. Sands, N. Sauer, and R. E. Woodrow, *Two-coloring all two-element maximal antichains*, J. Combin. Theory Ser. A, 57 (1991), pp. 109–116.

[5] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.

[6] M. Grötschel, L. Lovász, and A. Schrijver, *Polynomial algorithms for perfect graphs*, in Topics on Perfect Graphs, Ann. Discrete Math. 21, C. Berge and V. Chvátal, eds., North–Holland, Amsterdam, 1984, pp. 325–356.

[7] C. Hoàng and C. McDiarmid, *On the divisibility of graphs*, Discrete Mathematics, 242 (2002), pp. 145–156.

[8] T. Jensen and B. Toft, *Graph Coloring Problems*, Wiley-Intersci. Ser. Discrete Math. Optim., John Wiley, New York, 1995, p. 244.

[9] A. Kotlov, *personal communication.*

[10] J. Kratochvíl and Zs. Tuza, *On the complexity of bicoloring clique hypergraphs of graphs*, in Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, PA, 2000, pp. 40–41.

[11] L. Lovász, *Coverings and colorings of hypergraphs*, in Proceedings of the 4th Southeastern Conference on Combinatorics, Graph Theory, and Computing, Utilitas Mathematica Publishing, Winnipeg, 1973, pp. 3–12.

[12] F. Maffray and M. Preissmann, *On the NP-completeness of the k-colorability problem for triangle-free graphs*, Discrete Math., 162 (1996), pp. 313–317.

[13] F. Maffray and M. Preissmann, *Sequential colorings and perfect graphs*, Discrete Appl. Math., 94 (1999), pp. 287–296.

[14] F. Maffray and B. A. Reed, *A description of claw-free perfect graphs*, J. Combin. Theory Ser. B, 75 (1999), pp. 134–156.

[15] C. McDiarmid, *Hypergraph colouring and the Lovász local lemma*, Discrete Math., 167/168 (1997), pp. 481–486.

[16] B. Mohar and R. Škrekovski, *The Grötzsch theorem for the hypergraph of maximal cliques*, Electron. J. Combin., 6 (1999), #R26.

[17] J. Mycielski, *Sur le coloriage des graphes*, Colloq. Math., 3 (1955), pp. 161–162.

[18] K. R. Parthasaraty and G. Ravindra, *The strong perfect graph conjecture is true for $K_{1,3}$-free graphs*, J. Combin. Theory Ser. B, 21 (1976), pp. 212–223.

[19] K. R. Parthasaraty and G. Ravindra, *The validity of the strong perfect graph conjecture for $(K_4 - e)$-free graphs*, J. Combin. Theory Ser. B, 26 (1979), pp. 98–100.

[20] H. J. Prömel and A. Steger, *Almost all Berge graphs are perfect*, Comb. Probab. Comput., 1 (1992), pp. 53–79.

[21] T. J. Schaefer, *The complexity of satisfiability problems*, in Conference Record of the Tenth Annual ACM Symposium on Theory of Computing, ACM, New York, 1978, pp. 216–226.

[22] A. Tucker, *Coloring perfect $(K_4 - e)$-free graphs*, J. Combin. Theory Ser. B, 43 (1987), pp. 313–318.

[23] S. H. Whitesides, *A method for solving some graph recognition and optimization problems, with applications to perfect graphs*, in Topics on Perfect Graphs, Ann. Discrete Math. 21, C. Berge and V. Chvátal, eds., North–Holland, Amsterdam, 1984, pp. 281–297.

SIAM J. DISCRETE MATH.
Vol. 17, No. 3, pp. 377–383

NUMBER THEORETIC DESIGNS FOR DIRECTED REGULAR GRAPHS OF SMALL DIAMETER*

WILLIAM D. BANKS[†], ALESSANDRO CONFLITTI[‡], AND IGOR E. SHPARLINSKI[§]

Abstract. In 1989, F. R. K. Chung gave a construction for certain directed h-regular graphs of small diameter. Her construction is based on finite fields, and the upper bound on the diameter of these graphs is derived from bounds for certain very short character sums. Here we present two similar constructions that are based on properties of discrete logarithms and exponential functions in residue rings modulo a prime power. Accordingly, we use bounds for certain sums with additive and multiplicative characters to estimate the diameter of our graphs. We also give a third construction that avoids the use of bounds for exponential sums.

Key words. exponential sums, character sums, extremal problems, graphs

AMS subject classifications. 05C35, 11L07, 11L40

DOI. 10.1137/S0895480101396676

1. Introduction.

We recall that, in a directed graph G, the distance between two vertices is defined to be the length of the shortest directed path joining them, and the diameter $D(G)$ of G is defined to be the maximum distance over all possible pairs of vertices.

We say the a directed graph G is *h-regular* if the in-degree and the out-degree at every node is equal to h.

In many applications, such as in the design of communication networks, it is required that the underlying h-regular graphs have sufficiently many nodes, and it is desirable not only to keep h as small as possible (in order to reduce the complexity of the network), but also to minimize the diameter (so that information can be transmitted efficiently).

In [2], a construction of graphs with the above properties is proposed using finite fields of the form $\mathbb{F}_{q^n}$. Namely, for any prime q and any integer $n \geq 2$ with $q > (n-1)^2$, the construction produces q-regular graphs $G(q,n)$ with $q^n - 1$ nodes and with diameter

$$D\left(G(q,n)\right) \leq 2n + \frac{4n \log n}{\log q - 2\log(n-1)}. \tag{1}$$

In [11], a more flexible construction has been proposed that produces h-regular graphs for any $h \geq q^{1/2+\varepsilon}$, $\varepsilon > 0$.

The inequality (1) of [2] is based on bounds for very short character sums considered in [1, 7], while the result of [11] is based on bounds for even shorter sums in [10]. All of these estimates are derived from the celebrated Weil bound.

There are several other similar constructions and bounds for character sums; see [3, 9]. An alternative approach to bounding the diameter of $G(q,n)$, which in

*Received by the editors October 18, 2001; accepted for publication (in revised form) July 3, 2003; published electronically January 6, 2004.

http://www.siam.org/journals/sidma/17-3/39667.html

[†]Department of Mathematics, University of Missouri, Columbia, MO 65211 (bbanks@math.missouri.edu). The work of this author was supported in part by NSF grant DMS-0070628.

[‡]Dipartimento di Matematica, Università degli Studi di Roma "Tor Vergata," Via della Ricerca Scientifica, I-00133 Roma, Italy (conflitt@mat.uniroma2.it).

[§]Department of Computing, Macquarie University, Sydney, NSW 2109, Australia (igor@ics.mq.edu.au). The work of this author was supported in part by ARC grant A00000184.

some cases gives improved estimates, is described in [4, 5]. This method makes use of the Weil bound in a more direct way, but it applies only when q is extremely large relative to n.

In this paper, we show that similar constructions can be applied to the design of directed h-regular graphs with small diameter over the residue ring $\mathbb{Z}_{p^n}$, where p is an odd prime. One construction is an exact analogue of the construction of [2]. The other is an additive variant whose analogue over finite fields does not seem possible owing to a general lack of good bounds for short sums of additive characters with exponential functions. We also give a third construction, again over $\mathbb{Z}_{p^n}$, which has small diameter and small regularity for certain choices of the parameters. Our estimates for this last construction do not depend on bounds for exponential sums.

For any integer h, we denote by S_h the set consisting of the first h positive integers that are not divisible by p. In our first design (the multiplicative case), we specify vertices of our graph by elements of $\mathbb{Z}_{p^n}^*$, and then we select an integer h and connect vertices $u \to v$ if and only if $uv^{-1} \in S_h$. We denote the corresponding graph by $\mathcal{G}_\times(h,p,n)$.

Next, let ϑ be a fixed primitive root modulo p^n. For every element $a \in \mathbb{Z}_{p^n}^*$, we can define the discrete logarithm $\operatorname{ind} a$ uniquely by the conditions

$$\vartheta^{\operatorname{ind} a} \equiv a \pmod{p^n}, \qquad 0 \le \operatorname{ind} a < (p-1)p^{n-1}.$$

In our second design (the additive case), we specify vertices of our graph by elements of $\mathbb{Z}_{p^n}$ and connect vertices $u \to v$ if and only if $u - v \in \mathbb{Z}_{p^n}^*$ and $\operatorname{ind}(u-v) \in [1,h]$. We denote the corresponding graph by $\mathcal{G}_+(h,p,n)$

In our third design, we specify the vertices of our graph by elements of $\mathbb{Z}_m$, where m is any integer greater than 1, and connect vertices $u \to v$ if and only if the integer $u-v$, reduced modulo m, has precisely one nonzero digit when written in base g. We denote the corresponding graph by $\overline{\mathcal{G}}(m,g)$. For a wide range of parameters, these graphs have a smaller diameter than the corresponding graphs from [2] with the same number of nodes and the same regularity.

Throughout the paper, $\log z$ denotes the natural logarithm of z. For any integer m, we denote by $\mathbf{e}_m$ the additive character $\mathbf{e}_m(z) = \exp(2\pi i z/m)$. Constants in the "$O$" symbol depend only on p.

2. Preparations. Let X be the set of $(p-1)p^{n-1}$ multiplicative characters modulo p^n, and let $X^* \subset X$ be the subset of all nonprincipal characters.

We need the following well-known statements.

Lemma 1. *For any $z \in \mathbb{Z}_{p^n}^*$,*

$$\sum_{\chi \in X} \chi(z) = \begin{cases} (p-1)p^{n-1} & \text{if } z = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 2. *For any $z \in \mathbb{Z}_{p^n}$,*

$$\sum_{a=0}^{p^n-1} \mathbf{e}_{p^n}(az) = \begin{cases} p^n & \text{if } z = 0, \\ 0 & \text{otherwise.} \end{cases}$$

As we have already mentioned, our results are based on bounds for short character sums. The first one is essentially Exercise 8 in Chapter 9 of [6] (note that the largest element of S_h is $hp/(p-1) + O(1)$).

LEMMA 3. *Let $p \geq 3$ be a fixed prime number and let $\chi \in X^*$. For any integer h, $p^2 \leq h \leq (p-1)p^{n-1}$, we define $r = n \log p / \log h$ so that $h^r = p^n$. Then the bound*

$$\left| \sum_{x \in S_h} \chi(x) \right| = O\left(h^{1-\alpha/r^2}\right)$$

holds for some absolute constant $\alpha > 0$.

Our second result is a combination of Lemma 2 (for $r \leq 3/2$) and Theorem 4 (for $r > 3/2$) of [8].

LEMMA 4. *Let $p \geq 3$ be a fixed prime number, let ϑ be a primitive root modulo p^n, and suppose that* $\gcd(a,p) = 1$. *For any integer h, $2 \leq h \leq (p-1)p^{n-1}$, let $r = n \log p / \log h$ as before. Then the bound*

$$\left| \sum_{x=1}^{h} \mathbf{e}_{p^n}(a\vartheta^x) \right| = O\left(h^{1-\beta/r^2}\right)$$

holds for some absolute constant $\beta > 0$.

LEMMA 5. *Let $p \geq 3$ be a fixed prime number, let ϑ be a primitive root modulo p^k, and suppose that* $\gcd(a,p) = 1$. *Then*

$$\sum_{x=1}^{(p-1)p^{k-1}} \mathbf{e}_{p^k}(a\vartheta^x) = \begin{cases} -1 & \textit{if } k = 1, \\ 0 & \textit{if } k \geq 2. \end{cases}$$

Proof. We have

$$\sum_{x=1}^{(p-1)p^{k-1}} \mathbf{e}_{p^k}(a\vartheta^x) = \sum_{x \in \mathbb{Z}^*_{p^k}} \mathbf{e}_{p^k}(ax) = \sum_{x=1}^{p^k} \mathbf{e}_{p^k}(ax) - \sum_{x=1}^{p^{k-1}} \mathbf{e}_{p^k}(apx),$$

and the result follows from Lemma 2. □

3. Main results. We can now prove our main results.

THEOREM 6. *Let $p \geq 3$ be a fixed prime number. For any integer h in the range $p^2 \leq h < (p-1)p^{n-1}$, let $r = n \log p / \log h$. Then the bound*

$$D\left(\mathcal{G}_\times(h,p,n)\right) = O(r^3)$$

holds, provided that $r = o(n^{1/3})$.

Proof. Two vertices $u, v \in \mathbb{Z}^*_{p^n}$ are connected by a path of the form

$$u = w_0 \to w_1 \to \cdots \to w_d = v$$

if and only if

$$x_{i+1} = w_i / w_{i+1} \in S_h, \qquad 0 \leq i \leq d-1.$$

Thus, u is connected to v along such a path if only if there exist integers $x_1, \ldots, x_d \in S_h$ such that

$$v = u \prod_{j=1}^{d} x_j.$$

Therefore, to show that $D\left(\mathcal{G}_\times(h,p,n)\right) \le d$, it suffices to prove that every element $w \in \mathbb{Z}_{p^n}^*$ can be represented in the form

$$w = \prod_{j=1}^{d} x_j, \qquad x_1, \dots, x_d \in S_h. \tag{2}$$

By Lemma 1, the number T of solutions to (2) is given by

$$T = \frac{1}{(p-1)p^{n-1}} \sum_{x_1,\dots,x_d \in S_h} \sum_{\chi \in X} \chi\left(w^{-1} \prod_{k=1}^{d} x_k\right);$$

hence it is enough to show that $T > 0$ for every choice of w. Now, pulling off the contribution from the principal character, we have

$$T = \frac{h^d}{(p-1)p^{n-1}} + \frac{1}{(p-1)p^{n-1}} \sum_{\chi \in X^*} \chi(w^{-1}) \left(\sum_{x \in S_h} \chi(x)\right)^d.$$

By Lemma 3, we see that for some constant $C > 0$ (depending only on p),

$$\left|T - \frac{h^d}{(p-1)p^{n-1}}\right| < C^d h^{d-\alpha d/r^2} = C^d h^d p^{-\alpha n d/r^3};$$

hence T will be positive if

$$C^d p^{-\alpha n d/r^3} < \frac{1}{(p-1)p^{n-1}}.$$

This we can ensure by choosing

$$d = \left\lfloor \frac{r^3 \log p}{\alpha \log p - n^{-1} r^3 \log C} \right\rfloor + 1.$$

Consequently, if $r = o(n^{1/3})$ as $n \to \infty$, it follows that the diameter of $\mathcal{G}_\times(h,p,n)$ will be less than $2\alpha^{-1} r^3$ for sufficiently large n. □

THEOREM 7. *Let $p \ge 3$ be a fixed prime number. For any integer h in the range $2 \le h < (p-1)p^{n-1}$, let $r = n \log p / \log h$. Then the bound*

$$D\left(\mathcal{G}_+(h,p,n)\right) = O(r^3)$$

holds, provided that $r = o(n^{1/3})$.

Proof. Two vertices $u, v \in \mathbb{Z}_{p^n}$ are connected by a path

$$u = w_0 \to w_1 \to \cdots \to w_d = v$$

if and only if $w_i - w_{i+1} \in \mathbb{Z}_{p^n}^*$, $0 \le i \le d-1$, and

$$x_{i+1} = \operatorname{ind}(w_i - w_{i+1}) \in [1,h], \qquad 0 \le i \le d-1.$$

Thus, u is connected to v along such a path if only if there exist integers $x_1, \dots, x_d \in [1,h]$ such that

$$u = v + \sum_{j=1}^{d} \vartheta^{x_j}.$$

To show that $D\left(\mathcal{G}_+(h,p,n)\right) \le d$, it suffices to prove that every element $w \in \mathbb{Z}_{p^n}$ can be represented in the form

$$w = \sum_{j=1}^{d} \vartheta^{x_j}, \qquad x_1, \dots, x_d \in [1,h]. \tag{3}$$

By Lemma 2, the number T of solutions to (3) is given by

$$\begin{aligned} T &= \frac{1}{p^n} \sum_{x_1,\dots,x_d \in [1,h]} \sum_{b=0}^{p^n-1} \mathbf{e}_{p^n}\left(-bw + b\sum_{j=1}^{d} \vartheta^{x_j}\right) \\ &= \frac{1}{p^n} \sum_{b=0}^{p^n-1} \mathbf{e}_{p^n}(-bw) \left(\sum_{x\in[1,h]} \mathbf{e}_{p^n}(b\vartheta^x)\right)^d \\ &= \frac{h^d}{p^n} + \frac{1}{p^n} \sum_{b=1}^{p^n-1} \mathbf{e}_{p^n}(-bw) \left(\sum_{x\in[1,h]} \mathbf{e}_{p^n}(b\vartheta^x)\right)^d . \end{aligned}$$

To show that $T > 0$, it suffices to show that the summation on the right is less than h^d in absolute value. To do this, we collect terms with $\gcd(b,p^n) = p^{n-k}$, $k = 1, \dots, n$, which gives

$$\begin{aligned} \left| \sum_{b=1}^{p^n-1} \mathbf{e}_{p^n}(-bw) \left(\sum_{x\in[1,h]} \mathbf{e}_{p^n}(b\vartheta^x)\right)^d \right| &\le \sum_{k=1}^{n} \sum_{\substack{b=1 \\ \gcd(b,p^n)=p^{n-k}}}^{p^n-1} \left| \sum_{x\in[1,h]} \mathbf{e}_{p^n}(b\vartheta^x) \right|^d \\ &= \sum_{k=1}^{n} \sum_{\substack{a=1 \\ \gcd(a,p)=1}}^{p^k-1} \left| \sum_{x\in[1,h]} \mathbf{e}_{p^k}(a\vartheta^x) \right|^d . \end{aligned}$$

For $p^{k-1}(p-1) \ge h$, we apply Lemma 4 directly to obtain

$$\left| \sum_{x\in[1,h]} \mathbf{e}_{p^k}(a\vartheta^x) \right| \ll h^{1-\beta \log^2 h / k^2 \log^2 p} \le h^{1-\beta \log^2 h / n^2 \log^2 p} = hp^{-\beta n/r^3}.$$

For $p^{k-1}(p-1) < h$, write h in the form $h = p^{k-1}(p-1)i + j$ with $i \ge 1$ and $0 \le j \le p^{k-1}(p-1) - 1$. If $k \ge 2$, then we use Lemma 5 together with Lemma 4 to derive

$$\begin{aligned} \left| \sum_{x\in[1,h]} \mathbf{e}_{p^k}(a\vartheta^x) \right| &= \left| \sum_{\nu=0}^{i-1} \sum_{x=\nu p^{k-1}(p-1)+1}^{(\nu+1)p^{k-1}(p-1)} \mathbf{e}_{p^k}(a\vartheta^x) + \sum_{x=ip^{k-1}(p-1)+1}^{p^{k-1}(p-1)i+j} \mathbf{e}_{p^k}(a\vartheta^x) \right| \\ &= \left| \sum_{x=1}^{j} \mathbf{e}_{p^k}(a\vartheta^x) \right| \ll j^{1-\beta \log^2 j / k^2 \log^2 p} \\ &\ll h^{1-\beta \log^2 h / n^2 \log^2 p} \ll hp^{-\beta n/r^3}. \end{aligned}$$

For $k = 1$, using Lemma 5, we obtain

$$\left|\sum_{x\in[1,h]} \mathbf{e}_p(a\vartheta^x)\right| = \left|\sum_{\nu=0}^{i-1}\sum_{x=\nu(p-1)+1}^{(\nu+1)(p-1)} \mathbf{e}_p(a\vartheta^x) + \sum_{x=i(p-1)+1}^{(p-1)i+j} \mathbf{e}_p(a\vartheta^x)\right|$$
$$\le i + j < h/(p-1) + p \le 2h/p,$$

provided that h is sufficiently large. Consequently, for some constant $C > 0$ (depending only on p), we have

$$\begin{aligned}\sum_{k=1}^{n}\sum_{\substack{a=1\\ \gcd(a,p)=1}}^{p^k-1}\left|\sum_{x\in[1,h]} \mathbf{e}_{p^k}(a\vartheta^x)\right|^d &< (p-1)(2h/p)^d + C^d h^d p^{n-\beta nd/r^3}\\ &< 2(2h/3)^d + C^d h^d p^{n-\beta nd/r^3},\\ &< \frac{h^d}{2} + C^d h^d p^{n-\beta nd/r^3},\end{aligned}$$

provided that $d \ge 4$ and h is sufficiently large. Hence, T will be positive for large values of h if $d \ge 4$, and

$$C^d h^d p^{n-\beta nd/r^3} < \frac{h^d}{2},$$

which we can ensure by choosing

$$d = \left\lfloor \frac{r^3 \log p + n^{-1} r^3 \log 2}{\beta \log p - n^{-1} r^3 \log C}\right\rfloor + 4.$$

Consequently, if $r = o(n^{1/3})$ as $n \to \infty$, it follows that the diameter of $\mathcal{G}_+(h,p,n)$ will be less than $2\beta^{-1}r^3$ for sufficiently large n. □

THEOREM 8. *For any integer $m \ge 2$ and any base $g \ge 2$, $\overline{\mathcal{G}}(m,g)$ is regular of degree $h = (g-1)K$ and diameter $D\left(\overline{\mathcal{G}}(m,g)\right) = K$, where*

$$K = \left\lfloor \frac{\log(m-1)}{\log g}\right\rfloor + 1.$$

Proof. Two vertices $u, v \in \mathbb{Z}_m$ are connected by a path

$$u = w_0 \to w_1 \to \cdots \to w_d = v$$

if and only if $w_i - w_{i+1}$, reduced modulo m, has at most one nonzero digit when written in base g; that is, $w_i - w_{i+1} \equiv ag^j \pmod m$ with $1 \le a \le g-1$ and $0 \le j \le K-1$. Since every element $w \in \mathbb{Z}_m$ can be expressed in the form

$$w = \sum_{j=0}^{K-1} a_j g^j,$$

the diameter of $\overline{\mathcal{G}}(m,g)$ is $d = K$. Since every node u is connected only to elements of the form $u + ag^j$, we also see that $\overline{\mathcal{G}}(m,g)$ is regular of degree $h = (g-1)K$. □

In particular, taking m equal to a power q^n, if $g = \lfloor q^{1/2} \rfloor + 1$ and $q \geq (2n+1)^2$, then $h \leq q$ and $D\left(\overline{\mathcal{G}}(q^n, g)\right) \leq 2n+1$, which is stronger than the bound (1) implies for the graphs constructed in [2]. Moreover, one sees that for any $\varepsilon \geq 0$, there exists $A > 0$ such that for $g = \lfloor q^{2/(2+\varepsilon)} \rfloor + 1$, $q > n^A$, and sufficiently large n, our graphs have q^n nodes of degree $h < q$ and diameter at most $(1+\varepsilon)n$. Indeed, taking $A = 1 + 3/\varepsilon$, for these parameters we obtain $K \leq (1+\varepsilon/2)n + 1 \leq (1+\varepsilon)n < q^{\varepsilon/(2+\varepsilon)}$, provided that n is large enough.

We also note that the graphs $\overline{\mathcal{G}}(m, g)$ have an obvious algorithm for finding the shortest path between two nodes using only $O(K)$ arithmetic operations in $\mathbb{Z}_m$.

Acknowledgment. The first two authors would like to thank Macquarie University for its hospitality during the preparation of this paper.

REFERENCES

[1] L. Carlitz, *Distribution of primitive roots in a finite field*, Quart. J. Math., Oxford Ser. (2), 4 (1953), pp. 4–10.

[2] F. R. K. Chung, *Diameters and eigenvalues*, J. Amer. Math. Soc., 2 (1989), pp. 187–196.

[3] F. R. K. Chung, *Spectral Graph Theory*, CBMS Reg. Conf. Ser. Math. 92, AMS, Providence, RI, 1997.

[4] S. D. Cohen, *Polynomial factorization, graphs, designs and codes*, in Finite Fields: Theory, Applications, and Algorithms, Contemp. Math. 168, AMS, Providence, RI, 1994, pp. 23–32.

[5] S. D. Cohen, *Polynomial factorization and an application to regular directed graphs*, Finite Fields Appl., 4 (1998), pp. 316–346.

[6] A. A. Karatsuba, *Basic Analytic Number Theory*, Springer-Verlag, Berlin, 1993.

[7] N. M. Katz, *An estimate for character sums*, J. Amer. Math. Soc., 2 (1989), pp. 197–200.

[8] N. M. Korobov, *On the distribution of digits in periodic fractions*, Mat. Sb. (N.S.), 89 (1972), pp. 654–670 (in Russian).

[9] W. W.-C. Li, *Number Theory with Applications*, World Scientific, Singapore, 1996.

[10] G. I. Perel'muter and I. E. Shparlinski, *Distribution of primitive roots in finite fields*, Uspekhi Mat. Nauk, 45 (1990), pp. 185–186 (in Russian).

[11] I. E. Shparlinski, *On parameters of some graphs from finite fields*, European J. Combin., 14 (1993), pp. 589–591.

SIAM J. DISCRETE MATH.
Vol. 17, No. 3, pp. 384–402

IMPROVED APPROXIMATION ALGORITHMS FOR THE DEMAND ROUTING AND SLOTTING PROBLEM WITH UNIT DEMANDS ON RINGS*

CHRISTINE T. CHENG[†]

Abstract. In the *demand routing and slotting problem* on unit demands (unit-DRSP), we are given a set of unit demands on an n-node ring. Each demand, which is a $(source, destination)$ pair, must be routed clockwise or counterclockwise and assigned a slot so that no two routes that overlap occupy the same slot. The objective is to minimize the total number of slots used.

It is well known that unit-DRSP is NP-complete. The best deterministic approximation algorithm guarantees a solution that is $2 \times OPT$. A demand of unit-DRSP can be viewed as a chord on the ring. Let w denote the size of the largest set of demand chords that mutually cross in the interior of the ring. We present a simple approximation algorithm that uses at most $(2 - 1/\lceil w/2 \rceil) \times OPT$ slots in an n-node network; this is the first deterministic approximation algorithm that beats the factor of 2 for *all* values of OPT and therefore for all instances of the input.

If randomization is allowed, an algorithm by Kumar produces, with high probability, a solution that uses asymptotically $(1.5 + \frac{1}{2e} + o(1)) \times OPT$ slots. However, when OPT is not large enough, the factor can exceed 2. In this paper, we show how combining our algorithm with Kumar's yields a randomized approximation algorithm that has, with high probability, a constant factor of $2 - 1/\theta(\log n)$. While asymptotically it is not better than Kumar's, the approximation factor holds for *all* values of OPT.

Key words. bandwidth allocation problem, demand routing and slotting, SONET rings, WDM networks

AMS subject classifications. 68, 90

DOI. 10.1137/S0895480101386723

1. Introduction.

Among the most popular configurations of *synchronous optical networks*, or SONETs, are rings (i.e., cycles). In a SONET ring, nodes are connected by links made of optical fibers. Each link in the ring has the same capacity K and is divided into K *slots*, labeled from 1 to K, where each slot has size equal to a unit of ring capacity. To transmit a *unit* demand between two nodes, a route, whether clockwise or counterclockwise, is chosen and is assigned a slot number. One slot number is used because unit demands must use the same slot for each of the links traversed. To transmit a demand of size d units between two pairs of nodes, a route must again be chosen and d slot numbers are assigned for the route. Once a slot in a link is assigned to a demand, it is occupied. No other demands that go in either direction of the link can use the same slot. We note that assigning slots to routes is equivalent to assigning colors to the routes so that no two overlapping routes are assigned the same color.

The cost of constructing a SONET ring is an increasing function of the capacity of the ring. Thus, before purchasing or constructing a SONET ring, it is important to determine the minimum number of slots needed to transmit all the demands in the network. (In practice, the goal is to satisfy the *expected* demands of the network.)

*Received by the editors March 22, 2001; accepted for publication (in revised form) July 8, 2003; published electronically January 6, 2004. This work supported in part by ONR grant N00014-96-1-0829. Parts of this paper first appeared in Randomization, Approximation and Combinatorial Optimization, Lecture Notes in Comput. Sci. 1671, Springer-Verlag, New York, 1999, pp. 209–220. This is the proceedings of APPROX '99.

http://www.siam.org/journals/sidma/17-3/38672.html

[†]Department of EECS, University of Wisconsin-Milwaukee, 3200 N. Cramer St., Milwaukee, WI 53211 (ccheng@cs.uwm.edu).

The *demand routing and slotting problem* (DRSP) on rings is stated as follows: given a set of demands on the ring, what is the minimum number of slots needed so that (i) each demand is routed, (ii) each route assigned a set of slots, and (iii) no two routes that use the same edge of the ring are assigned the same slots?

Carpenter, Cosares, and Saniee [5] showed that DRSP on rings is NP-hard. They presented several approximation algorithms whose solutions are within a factor of 2 of the optimal. Sometimes, demands *can* be split, but only at integral values, and can be regarded as a multiplicity of unit demands [19]. Thus, an important subcase of the problem is when all the demands have unit size. Surprisingly, the complexity of the problem remains the same [6]. The best deterministic approximation algorithm (given first by Raghavan and Upfal in [18]) does not have a better approximation factor either. We shall refer to this subcase of DRSP as *unit-DRSP*.

In this paper, we present the first deterministic approximation algorithm for unit-DRSP that has an approximation factor less than 2 for *all* instances of the input. A unit of demand can be viewed as a chord on the ring whose endpoints correspond to the source and destination of the demand. Let w denote the size of a largest set of demand chords that mutually intersect at a point[1] in the interior of the ring. We show that it is always possible to route and slot all demands using at most $(2 - 1/\lceil w/2 \rceil) \times OPT$ slots in an n-node network in time $O(|I|n^2)$, where I is the set of all demands. Since demand chords are incident to two nodes of the ring, $w \leq \lfloor n/2 \rfloor$, and so in the worst case our algorithm achieves an approximation factor of $2 - 8/(2n + 4)$. Finally, we note that currently deployed SONET rings typically have at most 24 nodes [4]. In this case, our algorithm guarantees a solution that is at most $1.83 \times OPT$.

If we allow randomization, a recent Monte Carlo approximation algorithm by Kumar [14] has a factor of $1.5 + \frac{1}{2e} + o(1)$, where e is the base of the natural logarithm; i.e., asymptotically, the algorithm is a 1.68-randomized approximation algorithm. We note, however, that when OPT is not large enough, the $o(1)$ additive term can become significant so that the approximation factor exceeds 2. In this paper, we show how combining our algorithm with Kumar's produces a randomized approximation algorithm that has a constant factor of $2 - 1/\theta(\log n)$. Thus, while there are instances when Kumar's algorithm has a better guarantee, our approximation factor holds for *all* values of OPT and is an improvement over the $2 - 1/\theta(n)$ approximation factor of our deterministic algorithm.

1.1. Overview of the paper. The remainder of the paper is organized as follows. In section 2, we define terms, give two lower bounds that are relevant to our method, and present a 2-approximation algorithm for unit-DRSP. In section 3, we present a greedy algorithm for coloring circular arcs due to Tucker and discuss some of its properties. We then describe our deterministic algorithm in section 4 and our randomized algorithm in section 5.

1.2. Related work. We remark that unit-DRSP has also been studied extensively in SONETs and WDM (wavelength division multiplexing) networks with different topologies, including trees, trees of rings, and meshes [18, 6, 15, 16, 17, 2]. Beauquier et al. survey the most recent results in the area [1]. While the earliest model of the unit-DRSP in WDM ring networks just reduces to the problem above [18, 14], another model assumes that the underlying ring network is *directed* and *symmetric.*

[1] We emphasize that the intersection of two chords in this set must be just a *single* point in the interior of the ring. Thus, if two demand chords have the same endpoints, only one of them can be part of the set.

That is, if edge $(i, i+1) \in E(G)$, then $(i+1, i) \in E(G)$. Thus, two demands that traverse a link of the network, but in opposite directions, can be assigned the same slot number since they use, technically, different edges of the network. Wilfong and Winkler considered this problem and showed that unit-DRSP remains NP-hard in this setting [21]. The best approximation algorithm known for this problem also has a factor of 2.

2. Preliminaries.

The unit-DRSP problem is defined on an n-node ring. We are given a set of demands I on a ring network. Each demand is a (*source*, *destination*) pair where the source and destination are distinct nodes on the network. A *routing* for I is an assignment to each demand of either the clockwise (which, henceforth, we abbreviate as cw) or counterclockwise (ccw) source-destination path. A *slotting* for a routing of I is equivalent to assigning colors to the $|I|$ paths so that no two overlapping paths are assigned the same color. (Paths P_1 and P_2 are said to *overlap* if they have at least one edge of the network in common.) Since a path on the ring network can be viewed as an arc on a circle, we shall use the terms "paths" and "arcs" interchangeably.

A fixed choice of one of the cw or ccw paths for each demand in I determines a set of circular arcs C; conversely, a set of circular arcs C is *derivable* from I if the arcs are obtained from routing all the demands in I. We let $\mathcal{D}(I)$ denote the collection of arc-sets derivable from I. A solution of I consists of some $C \in \mathcal{D}(I)$ and a valid coloring of C. An optimal solution to unit-DRSP is one that uses the fewest number of colors among all possible solutions.

Let C be a set of circular arcs. A pair of paths in C forms a *conflicting pair* if the paths overlap and their union is the entire ring. Assume C is derivable from I. We say C is a *parallel routing* if C does not contain any conflicting pairs. We say that $A \subseteq C$ is an *independent set of arcs* if no two arcs in A overlap. Thus, arcs that are assigned the same color in a solution of unit-DRSP form an independent set.

2.1. Lower bounds.

We first establish some lower bounds on the number of colors required by an optimal solution to unit-DRSP in terms of properties of I and the routings in $\mathcal{D}(I)$. One way to view a demand in I is to consider it as a chord on the circle, where the endpoints of the chord correspond to the source and destination of the demand. Let us say that two demand chords *cross* each other if they intersect at a single point in the interior of the circle. It is easy to see that the paths of two demands that cross each other will always overlap and must be assigned different colors. Thus, the size of the largest set of demand chords that mutually cross each other is a natural lower bound to the number of colors needed for an optimal solution of unit-DRSP.

Let us denote by G_I the graph whose vertex set corresponds to the set of distinct demand chords generated from the demands of I. Two vertices in the graph are adjacent if and only if their corresponding demand chords cross. The parameter of interest to us is the clique number of G_I, $\omega(G_I)$. By the above discussion, $\omega(G_I)$ is a lower bound on the optimal number of colors for the unit-DRSP. Gavril showed that $\omega(G_I)$ can be computed in time $O(|I|^3)$ [9]. Bhattacharya and Kaller later improved the running time to $O(|I| + n \log n)$ [3].

Let e be an edge on the ring. A set of circular arcs C in $\mathcal{D}(\mathcal{I})$ induces a *load* L_e on e, where L_e is the number of arcs in C that contain e. Set C also induces a *ringload* of $L_C = \max_e\{L_e\}$ on the ring. It is not difficult to see that if we wish to assign colors to each arc in C so that overlapping arcs are assigned different colors, then at least L_C colors must be used.

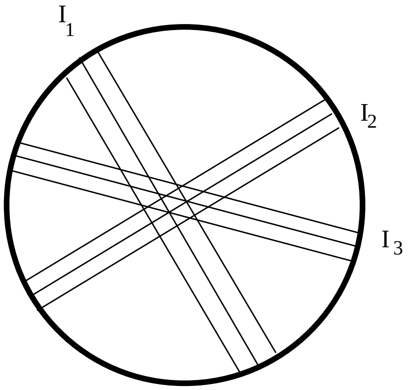


FIG. 2.1. *An example of what the demand chords of I could look like in our example if $k = 3$.*

Thus, another lower bound for unit-DRSP can be obtained by modifying the objective function to minimize the ringload of the network. The ringloading problem defined on I asks for C^* such that $L_{C^*} = \min_{C \in \mathcal{D}(I)} L_C$. Schrijver, Seymour, and Winkler [19] showed how to find C^* so that C^* is also a parallel routing in time $O(|I|n^2)$. Frank [7] also addressed this problem. Since an optimal solution to unit-DRSP must have a ringload of at least L_{C^*}, a coloring of the optimal solution uses at least L_{C^*} colors. We have thus shown the following.

PROPOSITION 2.1. *The optimal solution to unit-DRSP uses at least* $\max\{\omega(G_I), L_{C^*}\}$ *colors, where C^* is an optimal solution to the ringloading problem on I.*

We must point out, however, that our lower bound can be a weak one; i.e., it can be significantly smaller than the number of slots used by the optimal solution to unit-DRSP. Consider the following example, where the vertices of the ring are labeled from 1 to $6k$ in the cw direction and the set of demands $I = I_1 \cup I_2 \cdots \cup I_k$ is such that each $I_j = \{d_{j1} = (3(j-1)+1, 3k+3(j-1)+3), d_{j2} = (3(j-1)+2, 3k+3(j-1)+2), d_{j3} = (3(j-1)+3, 3k+3(j-1)+1)\}$. An interesting property of this I is that when $d \in I_i$ and $d' \in I_j$ then the demand chords of d and d' intersect in the interior of the circle if and only if $i \neq j$. (See Figure 2.1.)

Clearly, $\omega(G_I) = k$. If we route d_{j1} ccw and d_{j3} cw for each I_j, the resulting arcs induce a load of k on the ring. If we also route d_{j2} cw when j is odd and ccw when j is even, these additional arcs induce an additional load of $\lceil \frac{k+1}{2} \rceil$. So we have found a routing that induces a ringload of $k + \lceil \frac{k+1}{2} \rceil$ and thus, $L_{C^*} \leq k + \lceil \frac{k+1}{2} \rceil$. Thus, $\max\{\omega(G_I), L_{C^*}\} \leq \max\{k, k + \lceil \frac{k+1}{2} \rceil\}$.

Suppose $i \neq j$. All the chords of the demands of I_i intersect with all the chords of the demands of I_j. Thus, the slots occupied by the routed demands of I_i must be distinct from the slots occupied by the routed demands of I_j. Furthermore, the routed demands of each I_j must occupy at least two of these slots since two demands have to be routed in the same direction. Thus, the minimum number of slots needed by the optimal solution to unit-DRSP is at least $2k$, which is roughly $4/3 \max\{k, k + \lceil \frac{k+1}{2} \rceil\}$, provided $k \geq 2$.

2.2. A 2-approximation scheme. Recall that a solution for unit-DRSP consists of a routing $C \in \mathcal{D}(I)$ and a valid coloring of C. In the 2-approximation algorithm of unit-DRSP, an optimal solution to the ringloading problem on I, C^*, is the chosen routing. The arcs in C^* are colored as follows. Let $S_e \subseteq C^*$ be the set of arcs that pass through edge e. Note that $|S_e| \leq L_{C^*}$. All the arcs in S_e are assigned different colors since they overlap. On the other hand, the set of arcs that do not pass through e, $C^*\backslash S_e$, forms what is called an *interval set* since the arcs can be considered as intervals on a line. It is known that such a set of arcs can be colored optimally in linear time, where the number of colors used is $L_{C^*\backslash S_e}$ [10]. Thus, the total number of colors used is at most $|S_e| + L_{C^*\backslash S_e} \leq 2L_{C^*}$.

The 2-approximation algorithm above is an example of one basic approach to solving the unit-DRSP: first, route all demands, and then color the arcs. Our algorithm also initially routes all demands. What differentiates our algorithm from this basic approach is that whenever it detects suboptimality as it colors, it reroutes some of the demands. This crucial step allows us to show that the approximation factor of our algorithm is better than 2.

3. Coloring circular arcs. We examine the problem of coloring a fixed $C \in \mathcal{D}(I)$. Recall that a valid coloring for C assigns a color to each arc in C so that no two overlapping arcs have the same color. The problem of finding a valid coloring for C that uses as few colors as possible has been studied extensively. Garey et al. showed that the problem is NP-complete [8]. Tucker gave a simple 2-approximation scheme that used the ringload of C as a lower bound [20]. He also conjectured that $\frac{3}{2} \times \omega(C)$ colors will be sufficient to properly color C, where $\omega(C)$ is the size of the largest set of arcs that mutually overlap each other. In 1980, Karapetian proved that this conjecture is indeed true [12]. Independently, Hsu and Shih gave a 5/3-approximation scheme [11]. More recently, Kumar gave a randomized approximation algorithm that achieves a factor of $1 + 1/e + o(1)$ [13].

We present GREEDY, an algorithm for coloring the arcs of C due to Tucker [20] that will be used repeatedly as a subroutine in our new algorithm. For an arbitrary arc a, we shall call the first and second endpoints traversed in the cw direction its *left* (denoted by $l(a)$) and its *right* endpoint (denoted by $r(a)$), respectively.

Let $cmax = 1$. Start with an arbitrary arc a in C and color it with $cmax$. In the cw direction, we proceed to the next arc b, an arc whose left endpoint is closest to the right endpoint of a (in the cw direction) and color it with $cmax$ unless a and b overlap, in which case we color b with $cmax + 1$. We continue this process where the next arc chosen is always an uncolored arc whose left endpoint is closest to the right endpoint of the current arc in the cw direction. We color the next arc with $cmax$ whenever possible. Otherwise, we update $cmax$ with $cmax + 1$ and color with the new $cmax$. (See Figure 3.1 for the pseudocode and Figure 3.2 for an example.) We store all the arcs colored i in the set C_i and let $c_{i,j}$ denote the jth arc colored i. The last arc colored i will also be denoted as c_{i,n_i}.

Let us denote the left endpoint of arc a, the first arc we colored, as vertex 1. We then label the rest of the nodes in the ring from 2 to n starting from vertex 1 in the cw direction [2]. We say that the algorithm has completed k *rounds* if the algorithm has traversed vertex 1 $k+1$ times. We let R_k be the set of arcs colored during the kth round. If an arc's left endpoint was traversed during the kth round but its right endpoint during the $(k+1)$st round, we consider the arc to be in R_k.

LEMMA 3.1. *If, at some round,* GREEDY *used exactly* 2 *colors, say* $i-1$ *and* i, *then* $1 \leq l(c_{i-1,1}) < r(c_{i,1}) < l(c_{i,1}) \leq n$ *and* $c_{i,1}$ *was the last arc colored during the*

[2] We point out that this numbering scheme is not fixed for our Unit-DRSP Algorithm in the following sense. Our algorithm may call GREEDY more than once. For each call, the first colored arc is randomly picked. Hence, the ccw endpoint of the first colored arc of the ith call is vertex 1 for that instance, but it may have a different number for the jth call when $i \neq j$.

```
GREEDY($C, a$)
*(Preprocessing step: sort endpoints of arcs and store in a linked list. Let $next(a)$ be a pointer to the arc whose left endpoint is closest to the right endpoint of $a$ in the cw direction.)*
$cmax \leftarrow 1$, $j \leftarrow 1$;
$D_i \leftarrow \emptyset, i = 1, \ldots, |C|$;
$C \leftarrow C \setminus \{a\}$
while $next(a) \neq nil$ and $next(a)$ is not colored
    if $next(a)$ overlaps with an arc colored $cmax$
        then $n_{cmax} \leftarrow j$, $cmax$++, $j \leftarrow 1$          /* use a new color */
        else $j$++;                                  /* keep the same color */
    $c_{cmax,j} \leftarrow next(a)$;
    $a \leftarrow next(a)$;
    $C \leftarrow C \setminus \{a\}$;
for $i = 1$ to $c_{max}$
    $D_i \leftarrow \{c_{i,1}, \ldots c_{i,n_i}\}$
return($C, \{D_1, D_2, \ldots, D_{cmax}\}$);
```

FIG. 3.1. *The* GREEDY *algorithm for coloring circular arcs.*

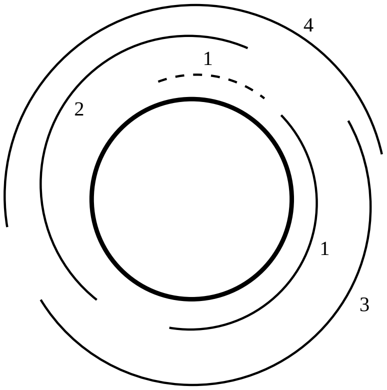


FIG. 3.2. *An example of how* GREEDY *labels the arcs if it started with the arc drawn with dashed lines.*

round.

Proof. Suppose at the beginning of a round, $cmax = i - 2$ and arcs $c_{i-1,1}$ and $c_{i,1}$ were both colored during the round. This means that $1 \leq l(c_{i-1,1}) < r(c_{i-1,1}) \leq l(c_{i,1}) \leq n$ and the endpoints of all the arcs in C_{i-1} lie between vertex 1 and $l(c_{i,1})$. Since $c_{i,1}$ must overlap with an arc colored $i - 1$, it must go beyond vertex 1 and $l(c_{i-1,1})$, so $r(c_{i,1}) < l(c_{i,1})$ and $l(c_{i-1,1}) < r(c_{i,1})$. Consequently, arc $c_{i,1}$ must be the last arc colored in the round. □

An immediate implication of this lemma is that *cmax* increases by at most two in each round of coloring.

THEOREM 3.2. GREEDY *uses at most two colors during each round of coloring.*

Suppose we run the GREEDY algorithm and stop it before all the arcs are colored. Below, we describe a relationship between the number of rounds GREEDY has completed and the load induced by the uncolored arcs. The proof of the theorem can be found in the appendix.

Theorem 3.3. *Let $0 \leq k \leq L_{C^*}$ and let f be an edge of the ring. If* GREEDY *traversed edge f k times, then the load induced by the remaining uncolored arcs on f is at most $L_{C^*} - k$.*

If GREEDY traversed the ring k times, then it must traverse each edge of the ring k times as well. Here is a direct consequence of the theorem.

Corollary 3.4. *Let $0 \leq k \leq L_{C^*}$ and let f be an edge of the ring. If* GREEDY *traversed the ring k times, then the load induced by the remaining uncolored arcs on f is at most $L_{C^*} - k$.*

4. Description and analysis of algorithm. To give the reader an idea of our algorithm, let us first consider the simple case when $\omega(G_I) \leq 3$. That is, at most three demand chords in I mutually cross each other. Our algorithm consists of two basic steps.

(a) Start with C^*, an optimal solution to the ringloading problem I, such that C^* is also a parallel routing. (We know that such a routing exists and can be found in $O(|I|n^2)$ time [19].)

(b) While some arc is not colored, apply GREEDY to C^* for two rounds. Remove the colored arcs.

From Theorem 3.2, GREEDY uses at most four colors for every two rounds of coloring. If it uses at most three colors, we do no modifications. However, if it uses four colors, we argue below that an arc can be rerouted so that this newly rerouted arc together with all the arcs discovered in two rounds of GREEDY can be colored with three colors. Hence, if our algorithm goes through $2k$ rounds of GREEDY, then it uses at most $3k$ colors.

Suppose at the end of two rounds of coloring, GREEDY used four colors; i.e., GREEDY used colors 1 and 2 during the first round and colors 3 and 4 during the second round. From the proof of Lemma 3.1, we know that $c_{2,1}$ is an arc that goes beyond vertex 1, so the endpoints of $c_{2,1}, c_{3,1}$, and $c_{4,1}$ arcs colored during the second round have the following ordering:

$$1 < r(c_{2,1}) \leq l(c_{3,1}) < r(c_{3,1}) \leq l(c_{4,1}) \leq n. \tag{4.1}$$

This means that all the endpoints of arcs in $C_2 \backslash \{c_{2,1}\}$ lie between $r(c_{2,1})$ and $l(c_{3,1})$. Since $c_{3,1}$ must overlap with some arc in C_2, it must be the case that

$$l(c_{3,1}) \leq l(c_{2,1}) < r(c_{3,1}). \tag{4.2}$$

From Lemma 3.1, we also have

$$1 \leq l(c_{3,1}) < r(c_{4,1}) < l(c_{4,1}) \leq n. \tag{4.3}$$

Inequalities (4.2) and (4.3) together with the assumption that C^* has no conflicting pairs imply a stronger version of (4.1).

$$1 < r(c_{2,1}) < l(c_{3,1}) < r(c_{3,1}) < l(c_{4,1}) \leq n. \tag{4.4}$$

Combining (4.4) and (4.3), and (4.4) and (4.2), we have

$$1 < r(c_{2,1}) < l(c_{3,1}) < r(c_{4,1}), \tag{4.5}$$

$$l(c_{2,1}) < r(c_{3,1}) < l(c_{4,1}) \leq n. \tag{4.6}$$

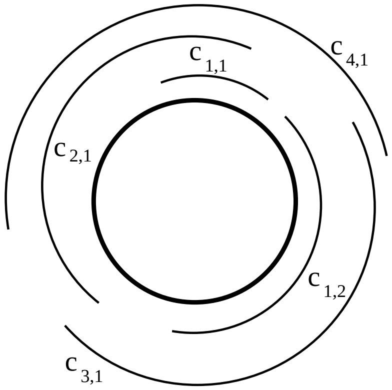


FIG. 4.1. *In this example, $c_{1,*}$ is $c_{1,1}$. Notice that the demand chords of $c_{2,1}, c_{3,1}$, and $c_{4,1}$ mutually cross each other, but $c_{1,*}$'s demand chord does not cross that of $c_{4,1}$. Thus, we can reroute $c_{4,1}$ so that now we have three independent sets, $\{c_{1,1}, \widehat{c_{4,1}}\}, \{c_{1,2}, c_{2,1}\}$, and $\{c_{3,1}\}$.*

And since C^* has no conflicting pairs, $r(c_{4,1}) < l(c_{2,1})$. Hence,

$$1 \le r(c_{2,1}) < l(c_{3,1}) < r(c_{4,1}) < l(c_{2,1}) < r(c_{3,1}) < l(c_{4,1}) \le n. \tag{4.7}$$

That is, the demand chords of $c_{2,1}, c_{3,1}$, and $c_{4,1}$ form a K_3 in G_I.

Let $c_{1,*}$ be the last arc in C_1 that overlaps nontrivially with $c_{2,1}$; i.e., $l(c_{1,*}) < r(c_{2,1}) \le l(g)$, where $g = c_{1,*+1}$ if $* \ne n_1$ and $g = c_{2,1}$ otherwise. Thus, either $l(c_{1,*}) < r(c_{2,1}) \le r(c_{1,*})$ or $l(c_{1,*}) < r(c_{1,*}) < r(c_{2,1})$. In both cases, $r(c_{1,*}) < l(c_{2,1})$ or else $c_{1,*}$ and $c_{2,1}$ form a conflicting pair. If, in addition, $r(c_{4,1}) < r(c_{1,*})$, then we have the inequalities

$$l(c_{1,*}) < r(c_{2,1}) < l(c_{3,1}) < r(c_{4,1}) < r(c_{1,*}) < l(c_{2,1}) < r(c_{3,1}) < l(c_{4,1}), \tag{4.8}$$

which imply the existence of a K_4 in G_I.

Since $\omega(G_I) \le 3$, it must be the case that $r(c_{4,1}) \ge r(c_{1,*})$. In other words, $1 \le l(c_{1,*}) < r(c_{1,*}) \le r(c_{4,1}) < l(c_{4,1}) \le n$. If we reroute $c_{4,1}$ so that it no longer passes through vertex 1, then it will not overlap any of the arcs $c_{1,1}, c_{1,2} \dots c_{1,*}$. Let $\widehat{c_{4,1}}$ denote the rerouted $c_{4,1}$. The set $\{\widehat{c_{4,1}}\} \cup \{c_{1,1}, c_{1,2} \dots c_{1,*}\}$ is an independent set.

Since GREEDY colored the arcs $\{c_{2,1}\} \cup C_1 \backslash \{c_{1,1}, c_{1,2} \dots c_{1,*}\}$ in the first round and since $c_{1,*}$ was the last arc in C_1 that overlapped with $c_{2,1}$, the set $\{c_{2,1}\} \cup C_1 \backslash \{c_{1,1}, c_{1,2} \dots c_{1,*}\}$ is an independent set too. Similarly, since GREEDY colored the arcs in $C_3 \cup C_2 \backslash \{c_{2,1}\}$ during the second round and $c_{4,1}$ was the last arc colored during the round, $l(c_{2,2}) > 1$ and $r(c_{3,n_3}) \le n$. That is, none of the arcs in C_3 overlapped with the arcs in $C_2 \backslash \{c_{2,1}\}$. Thus, $C_3 \cup C_2 \backslash \{c_{2,1}\}$ is also an independent set.

Let us summarize what we have learned. If GREEDY used four different colors in two rounds of coloring, then the demands chords discovered by GREEDY during these two rounds must have K_3 as a subgraph in G_I. Since $\omega(G_I) \le 3$, we can reroute $c_{4,1}$ and rearrange the sets so that now $R_1 \cup R_2$ can be partitioned into three independent sets (instead of four): $\{\widehat{c_{4,1}}\} \cup \{c_{1,1}, c_{1,2} \dots c_{1,*}\}$, $\{c_{2,1}\} \cup C_1 \backslash \{c_{1,1}, c_{1,2} \dots c_{1,*}\}$, and $C_3 \cup C_2 \backslash \{c_{2,1}\}$. That is, all the arcs discovered by GREEDY in two rounds can be colored with three colors. These are precisely the steps our algorithm will take whenever it discovers that four colors were used at the end of two rounds of coloring. See Figure 4.1 for an example.

LEMMA 4.1. *Suppose $R_2 \cap C_4 \neq \emptyset$. The demand chords of $c_{2,1}, c_{3,1}$, and $c_{4,1}$ form a K_3 in G_I. If, in addition, $\omega(G_I) \leq 3$, then the arc $c_{4,1}$ can be rerouted so that only three colors are used to color the demands encountered in rounds* 1 *and* 2.

From Corollary 3.4, we know that whenever GREEDY traverses the ring k times, the remaining uncolored arcs induce a load of at most $L_{C^*} - k$ on the edges of the ring. This means that our algorithm can go through at most $2(L_{C^*}/2)$ rounds of GREEDY if L_{C^*} is even, and at most $2(\frac{L_{C^*}+1}{2})$ rounds of GREEDY if L_{C^*} is odd. We have the following result.

THEOREM 4.2. *Let I be a set of demands on the ring such that $\omega(G_I) \leq 3$. The* Unit-DRSP Algorithm *produces a routing and a slot assignment of the demands in I that uses at most* $1.5 \times OPT$ *slots.*

Proof. If L_{C^*} is even, the algorithm goes through at most L_{C^*} rounds, so it uses at most $\frac{3}{2} * L_{C^*}$ colors. If L_{C^*} is odd, and the algorithm goes through at most $L_{C^*} - 1$ rounds, it also uses at most $\frac{3}{2} * L_{C^*}$ colors. However, suppose, after $L_{C^*} - 1$ rounds, some arcs are still not colored. From Theorem 3.3, the remaining uncolored arcs induce a load of 1 on the edges of the ring. Thus, none of the arcs overlap and one color will be sufficient to color all of them. The total number of colors used then is at most $3(L_{C^*} - 1)/2 + 1 < \frac{3}{2} * L_{C^*}$. Since $L_{C^*} \leq OPT$, our theorem follows. □

Let us now consider the general case where $\omega(G_I) \leq 2z - 1$, where z is a positive integer. Our algorithm's two basic steps are as follows.

(a) Compute C^*, where C^* is a parallel routing.

(b) While some arc is not colored, apply GREEDY to C^* for z rounds. Remove the colored arcs.

Again, GREEDY can only use at most $2z$ colors for every z rounds of coloring. We extend Lemma 4.1 to show that because $\omega(G_I) \leq 2z - 1$, $2z - 1$ colors will be sufficient to slot all the demands encountered in the z rounds of coloring.

LEMMA 4.3. *If $R_k \cap C_{2k} \neq \emptyset$, then the following sets of inequalities are true.*

(i) $l(c_{1,*}) < r(c_{2,1}) < l(c_{3,1}) < \cdots < l(c_{2i-1,1}) < r(c_{2i,1}) < \cdots < r(c_{2k,1})$ *and*

(ii) $r(c_{1,*}) < l(c_{2,1}) < r(c_{3,1}) < \cdots < r(c_{2i-1,1}) < l(c_{2i,1}) < \cdots < l(c_{2k,1})$,

where $c_{1,*}$ is the last arc in C_1 that overlaps nontrivially with $c_{2,1}$ during the implementation of the GREEDY *algorithm.*

Proof. First, suppose $k = 1$. If $R_1 \cap C_2 \neq \emptyset$, then, from Lemma 3.1, $R_1 = C_1 \cup \{c_{2,1}\}$. By the definition of $c_{1,*}$, $1 \leq l(c_{1,*}) < r(c_{2,1})$ and since $r(c_{1,n_1}) \leq l(c_{2,1}) \leq n$, we have $r(c_{1,*}) \leq l(c_{2,1})$. Note, however, that $r(c_{1,*}) \neq l(c_{2,1})$ since equality would imply that $c_{1,*}$ and $c_{2,1}$ form a conflicting pair. Thus, the inequalities in (i) and (ii) are true when $k = 1$.

Suppose that after $i > 1$ rounds $R_i \cap C_{2i} \neq \emptyset$ and that our lemma holds for $k \leq i - 1$. From Lemma 3.1, we know that arcs $c_{2i-2,1}$ and $c_{2i,1}$ were the last arcs colored in rounds $i - 1$ and i, respectively. Using the same argument we made to arrive at inequalities (4.5) and (4.6), where we substitute $c_{2i-2,1}$ for $c_{2,1}$, $c_{2i-1,1}$ for $c_{3,1}$, and $c_{2i,1}$ for $c_{4,1}$, we have the following inequalities:

$$1 < r(c_{2i-2,1}) < l(c_{2i-1,1}) < r(c_{2i,1}), \tag{4.9}$$

$$l(c_{2i-2,1}) < r(c_{2i-1,1}) < l(c_{2i,1}) \leq n. \tag{4.10}$$

These inequalities, together with our assumption that the inequalities in (i) and (ii) hold for $k \leq i - 1$, show that our lemma holds for $k \leq i$. By induction, the lemma holds in general. □

In view of the lower bound $\omega(G_I)$, the next theorem tells us that if GREEDY traverses k rounds of coloring and uses the maximum possible number of $2k$ colors, then the number of colors used for the arcs encountered in the k rounds is within one of optimal. Furthermore, if the number of colors used is not optimal, we can reroute some arc so that there is no need to use the extra color.

THEOREM 4.4. *Let $k \geq 2$. If $R_k \cap C_{2k} \neq \emptyset$, then the demand chords of $c_{2,1}, c_{3,1}, \ldots, c_{2k,1}$ form a K_{2k-1} in G_I. In addition, if the demand chords of $c_{1,*}, c_{2,1}, c_{3,1}, \ldots, c_{2k,1}$ do not form a K_{2k} in G_I, then by rerouting $c_{2k,1}$ we can color the arcs with at most $2k-1$ colors.*

Proof. From the previous lemma, $r(c_{2,1}) < r(c_{2k,1})$ and $l(c_{2,1}) < l(c_{2k,1})$. If we also have $r(c_{2k,1}) \geq l(c_{2,1})$, then $c_{2,1}$ and $c_{2k,1}$ would form a conflicting pair. Thus, $r(c_{2k,1}) < l(c_{2,1})$. Combining the fact that $R_k \cap C_{2k} \neq \emptyset$ and inequalities (i) and (ii) of the previous lemma, we have the following inequalities:

$$r(c_{2,1}) < \cdots < l(c_{2k-1,1}) < r(c_{2k,1}) < l(c_{2,1}) < \cdots < r(c_{2k-1,1}) < l(c_{2k,1}). \tag{4.11}$$

This implies that the demand chords of arcs $c_{2,1}, c_{3,1}, \ldots, c_{2k,1}$ mutually intersect each other and consequently form a K_{2k-1} in G_I.

Suppose we also know that the demand chords associated with the arcs in the set $\{c_{1,*}, c_{2,1}, c_{3,1}, \ldots, c_{2k,1}\}$ do not form a K_{2k} in G_I. Since the previous lemma holds, it follows that $r(c_{2k,1}) \geq r(c_{1,*})$. That is, the demand chords of $c_{1,*}$ and $c_{2k,1}$ do not intersect in the interior of the circle. Let us reroute $c_{2k,1}$ and denote the new arc as $\widehat{c_{2k,1}}$. Thus, $\{c_{1,1}, c_{1,2} \ldots c_{1,*}\} \cup \{\widehat{c_{2k,1}}\}$ is an independent set. Since $k \geq 2$, we use the same argument as in the proof of Lemma 4.1 to prove that the sets $C_1 \backslash \{c_{1,1}, c_{1,2} \ldots c_{1,*}\} \cup \{c_{2,1}\}$ and $C_2 \backslash \{c_{2,1}\} \cup C_3$ are independent. Finally, by construction, the sets C_4, $C_5, \ldots, C_{2k-1}$ must also be independent. If we assign all arcs belonging to the same independent set the same color, then we have used exactly $2k-1$ colors. □

We now describe the algorithm in Figure 4.2, which is based on the proof of the theorem above.

```
Unit-DRSP Algorithm ($\boldsymbol{I}$)

Compute for $C^*$ and $\omega(G_I)$.
Compute for the smallest integer $z$ s.t. $\omega(G_I) \leq 2z-1$.
    $\mathcal{Q} \leftarrow \emptyset$.
    while ($C^* \neq \emptyset$)
        $C_i \leftarrow \emptyset, i = 1, \ldots, 2z$;
        Choose $a \in C^*$ and apply GREEDY for $z$ rounds;
        if $C_{2z} \neq \emptyset$
            $c_{2z,1} \leftarrow$ rerouted $c_{2z,1}$
            $C_3 \leftarrow C_3 \cup C_2 \backslash \{c_{2,1}\}$
            $C_2 \leftarrow \{c_{2,1}\} \cup C_1 \backslash \{c_{1,1}, c_{1,2} \ldots c_{1,*}\}$
            $C_1 \leftarrow \{c_{1,1}, c_{1,2} \ldots c_{1,*}\} \cup \{c_{2z,1}\}$
            $C_{2z} \leftarrow \emptyset$
        $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{C_1, \ldots, C_{2z-1}\}$
        $C^* \leftarrow C^* \backslash \{C_1, \ldots, C_{2z-1}\}$
return($\mathcal{Q}$);
```

FIG. 4.2. *Our approximation algorithm for routing and slotting unit demands in rings.*

THEOREM 4.5. *Let $\omega(G_I) \leq 2z - 1$, where z is a positive integer. The* Unit-DRSP Algorithm *produces a routing and a slot assignment that uses at most* $(2 - 1/\lceil\omega(G_I)/2\rceil) \times OPT$ *slots.*

Proof. We first note that $z \leq \lceil\omega(G_I)/2\rceil$. Suppose our algorithm went through zk rounds of GREEDY. From Theorem 4.4, it used at most $(2z-1)k$ colors. If $zk \leq L_{C^*}$, then $k \leq L_{C^*}/z$, so the algorithm used at most $(2z-1)(L_{C^*}/z) \leq (2 - 1/\lceil\omega(G_I)/2\rceil) \times L_{C^*}$ colors.

But suppose $zk > L_{C^*}$. This means that after $z(k-1)$ rounds of GREEDY, there were still uncolored arcs. From Theorem 3.3, these remaining arcs induce a load of at most $q = L_{C^*} - z(k-1)$. We claim that $2q-1$ colors would suffice to color them. If not, from Theorem 4.4, $c_{2,1}, \ldots, c_{2q,1}$ form a K_{2q-1} in G_I. In the proof of that theorem, we showed that $1 < r(c_{2,1}) < r(c_{2q,1}) < n$, which means that $c_{2q,1}$ went through edge $(1,2)$. That is, GREEDY went past edge $(1,2)$ $q+1$ times before all the arcs were colored. According to Theorem 3.3, this means that the load induced by the remaining arcs was at least $q+1$, contradicting our assumption.

Thus, the number of colors used by GREEDY was at most $(2z-1)(k-1)+2q-1$. Using the fact that $q = L_{C^*} - z(k-1)$ and $zk > L_{C^*}$, we have

$$\begin{aligned}(2z-1)(k-1)+2q-1 &= 2L_{C^*} - k\\ &< 2L_{C^*} - L_{C^*}/z\\ &= (2-1/z)L_{C^*}\\ &\leq (2 - 1/\lceil\omega(G_I)/2\rceil) \times L_{C^*}. \qquad \square\end{aligned}$$

COROLLARY 4.6. *We have an $O(|I|n^2)$-time algorithm for the unit-DRSP on rings that produces a $(2-8/(2n+4))$-approximation to the optimal solution whenever $n \geq 3$.*

Proof. Schrijver, Seymour, and Winkler [19] proved that an optimal solution to the ringloading problem of I, C^* that is also a parallel-routing can be determined in $O(|I|n^2)$ time. Bhattacharya and Kaller [3] showed that it takes $O(|I| + n\log n)$ time to find a largest clique in G_I. Let $\mathcal{R}$ denote the set of arcs discovered during z rounds of GREEDY. It is easy to check that rerouting and coloring takes at most $O(|\mathcal{R}|)$ time. Thus, coloring and rerouting arcs in C^* takes $O(|I|)$ time. The bottleneck of the algorithm therefore is finding C^*. Finally, since each demand chord is incident to two nodes of the network, $\omega(G_I) \leq \lfloor n/2\rfloor$, so $\lceil\omega(G_I)/2\rceil \leq n/4 + 1/2$. The approximation factor follows from the previous theorem. $\square$

5. The hybrid randomized algorithm. In [14], Kumar presented an algorithm that considers $n-1$ binary integer programs. Let e be an edge of the ring. Each integer program is a formulation of the unit-DRSP problem with the added assumption that e and another edge of the ring, $f(\neq e)$, are *complements* of each other (see the appendix for a definition). There are $n-1$ binary integer programs because all possible candidates for f are considered. If z_i is the objective value of the ith integer program, then $OPT = \min_i z_i$. To approximate OPT, Kumar's algorithm relaxes each program and applies randomized rounding. Let z^*_{frac} be the best optimal fractional solution among the $n-1$ programs. Kumar showed that, with high probability (w.h.p.), the number of slots used by the algorithm is at most $(1.5 + 1/(2e) + f(n, z^*_{frac}))z^*_{frac}$ (e, in this case, is the base of the natural logarithm), where

$$(5.1)\; f(n, z^*_{frac}) = 2\sqrt{2(\ln n + 2\ln z^*_{frac})/z^*_{frac}} + \sqrt{2\ln z^*_{frac}/z^*_{frac}} + 8/(z^*_{frac})^2.$$

We observe that when n is large compared to z^*_{frac}, then the bound for the number of slots used in Kumar's algorithm can exceed $2 \times z^*_{frac}$. On the other hand, when n is fixed and $z^*_{frac} \to \infty$, the bound converges to $(1.5 + 1/(2e))z^*_{frac}$.

Let us combine the Unit-DRSP Algorithm with Kumar's algorithm. That is, given a set of unit demands I, route I using Kumar's algorithm and the Unit-DRSP Algorithm. Output the routing which uses the fewest slots. We call this combination the hybrid randomized algorithm (HRA). How many slots does this algorithm use? To bound the number of slots used by HRA, we need the following lemma whose proof can be found in the appendix.

LEMMA 5.1. $\omega(G_I) \leq z^*_{frac}$.

This means that when Kumar's algorithm has, w.h.p., an approximation factor of $1.5 + \frac{1}{2e} + f(n, z^*_{frac})$, Unit-DRSP Algorithm's approximation factor is at most $2 - 1/\lceil z^*_{frac}/2 \rceil$.

Let $z_0(n) = M \ln n$, where M is a large real number. When $z^*_{frac} \leq z_0(n)$, the Unit-DRSP Algorithm will use at most $(2 - 1/\lceil z_0(n)/2 \rceil) \times OPT$ slots since the algorithm's approximation factor is monotonically increasing. When $z^*_{frac} > z_0(n)$, w.h.p., Kumar's algorithm will use at most $(1.5 + 1/(2e) + f(n, z_0(n)))z^*_{frac}$ slots since this algorithm's approximation factor is monotonically decreasing with respect to z^*_{frac}. In the appendix, we prove that if we assume $n \geq 2$ (which is valid because a ring has to have at least two nodes) and we choose M appropriately, then $(1.5 + 1/(2e) + f(n, z_0(n))) \leq 2 - 1/\lceil z_0(n)/2 \rceil$; i.e., when $z^*_{frac} = z_0(n)$, w.h.p., Kumar's algorithm's approximation factor is not more than that of Unit-DRSP. Hence, w.h.p., HRA uses at most $(2 - 1/O(\ln n)) \times OPT$ slots.

THEOREM 5.2. *With high probability, the hybrid randomized algorithm above produces a solution for the unit-DRSP in rings that uses at most* $(2-1/O(\ln n)) \times OPT$ *slots.*

6. Future directions.

In section 3, we mentioned several ways of coloring circular arcs. Let $C \in \mathcal{D}(I)$. If we use any of these coloring algorithms, we have a solution to our problem. How close is the number of colors used in the solution to optimality?

Tucker's coloring scheme gives a 2-approximation based on the ringload of C. Thus, if this coloring is applied to $C^* \in \mathcal{D}(I)$, where C^* is a routing of I that minimizes ringload, then the number of colors used is at most $2L_{C^*}$.

Let G_C be the graph whose vertices correspond to the arcs in C and the edges correspond to the arcs that intersect in C. Karapetian's coloring scheme gives a 3/2-approximation scheme based on $\omega(G_C)$, the largest clique in G_C. This suggests a very interesting problem: *How do we route I so that the largest set of arcs that mutually overlap is as small as possible?* If we can answer this question optimally, then we have a 3/2-approximation algorithm to unit-DRSP.

7. Appendix.

7.1. Reduction of ringload by GREEDY.

Let us now prove Theorem 3.3.

THEOREM 3.3 *Let $0 \leq k \leq L_{C^*}$ and let f be an edge of the ring. If* GREEDY *traversed edge f k times, then the load induced by the remaining uncolored arcs on f is at most $L_{C^*} - k$.*

Proof. Throughout this proof, we remind the reader that if GREEDY has finished k rounds of coloring, then it has also traversed each edge of the ring at least k times. Futhermore, if the last arc colored by GREEDY went beyond vertex 1, then some edges of the ring were traversed $k + 1$ times.

Let $f = (u, u+1)$, where $1 \leq u \leq n$ and where vertex $n+1$ is the same as vertex 1. Let $\{d_1, d_2, \ldots, d_l\}$ be the maximal set of arcs GREEDY discovered as it traversed f the kth time. We note that d_l either went through edge f or did not. If the former is true, then GREEDY was coloring d_l as it traversed f; otherwise, d_l was the last arc GREEDY colored and none of the remaining uncolored arcs have a left endpoint in the set $\{r(d_l), r(d_l)+1, \ldots, l(f)-1, l(f)\}$.

Let $S_f^{(k)} \subseteq C^* \backslash \{d_1, \ldots, d_l\}$ denote the set of uncolored arcs that traverse edge f after GREEDY has traversed the edge k times. Thus, the load induced by $C^* \backslash \{d_1, \ldots, d_l\}$ on f is exactly $|S_f^{(k)}|$. To prove the theorem, we shall do a double induction, first on k and then on u, the left endpoint of edge f.

When $k = 0$, the theorem is true trivially. Suppose that for some value $k' \leq L_{C^*}$ and for all k satisfying $0 \leq k < k'$, the theorem holds for all the edges in the ring. We claim that the theorem still holds when the number of times GREEDY traverses the edge f is k'. Let us start with edge $f = (1, 2)$ and show that $|S_f^{(k')}| \leq L_{C^*} - k'$.

Since GREEDY has just finished $k' - 1$ rounds of coloring, from our induction hypothesis, $|S_f^{(k')}| \leq L_{C^*} - (k'-1)$. If $|S_f^{(k')}| \leq L_{C^*} - k'$, our claim is true. However, suppose $|S_f^{(k')}| = L_{C^*} - (k'-1)$. This implies that neither the last arc colored in round $k'-1$ nor the first arc in round k' traversed edge $(1, 2)$, because the load induced on edge $(1, 2)$ did not drop since GREEDY traversed it the $(k'-1)$st time. Consequently, none of the arcs in $S_f^{(k')}$ have a ccw endpoint at vertex 1.

Let us denote by b the arc in $S_f^{(k')}$ whose left endpoint is closest to vertex 1 in the cw direction. Thus, all the arcs in $S_f^{(k')}$ go through the edges $(l(b), l(b)+1), (l(b)+1, l(b)+2), \ldots, (n, 1), (1, 2)$. Since none of these arcs were colored during round $k'-1$, $S_f^{(k')} \subseteq S_{(l(b),l(b)+1)}^{(k')} \subseteq S_{(l(b),l(b)+1)}^{(k'-1)}$.

Since b was not colored during round $k'-1$, there must be an arc c in $R_{k'-1}$ that went through edge $(l(b), l(b)+1)$; otherwise, GREEDY would have colored arc b in this round. Since c was colored during round $k'-1$, $c \notin S_f^{(k')}$. Thus, $S_f^{(k')} \cup \{c\} \subseteq S_{(l(b),l(b)+1)}^{(k'-1)}$, so $|S_f^{(k')}| + 1 \leq |S_{(l(b),l(b)+1)}^{(k'-1)}|$. In other words, $L_{C^*} - (k'-1) + 1 \leq |S_{(l(b),l(b)+1)}^{(k'-1)}|$. This violates our induction hypothesis that the theorem holds for $k \leq k'-1$, so our claim must be true for $f = (1, 2)$.

Let us now suppose that not only does our theorem hold for all edges when k satisfies $0 \leq k < k'$ but it also holds for edges $(u, u+1), 1 \leq u < u' \leq n$, when $k = k'$. We shall show that the theorem is true for edge $f = (u', u'+1)$ when $k = k'$. From the induction hypothesis, $|S_f^{(k')}| \leq |S_f^{(k'-1)}| \leq L_{C^*} - (k'-1)$. Suppose $|S_f^{(k')}| = L_{C^*} - (k'-1)$. None of the arcs colored during the k'th round traversed edge $(u', u'+1)$; otherwise, $|S_f^{(k')}| = |S_f^{(k'-1)}| - 1 = L_{C^*} - k'$. Consequently, none of the arcs in $S_f^{(k')}$ start at u'. Thus, all the arcs in $S_f^{(k')}$ traverse edge $(u'-1, u')$. That is, $S_f^{(k')} \subseteq S_{(u'-1,u')}^{(k')}$, so $L_{C^*} - (k'-1) \leq |S_{(u'-1,u')}^{(k')}|$, violating our induction hypothesis. It must be the case that $|S_f^{(k')}| \leq L_{C^*} - k'$. By induction, our theorem is true. ☐

7.2. Kumar's algorithm for unit-DRSP. In [14], Kumar presented an approximation algorithm for unit-DRSP. The solution the algorithm outputs is based on the optimal fractional solution of a relaxed binary integer program, which we denote as z^*_{frac}. Our goal is to show that z^*_{frac} is bounded below by $\omega(G_I)$. Let us first discuss some observations Kumar made, which we state in the form of lemmas. Let I

be a set of demands on the ring.

Let us define a (c, w)-*color partition* of a family of circular arcs C to be a partition of C into c families of arcs $C_1, C_2, \ldots, C_c$ all with ringload 1 and a set of interval arcs S with ringload w. (That is, there is at least one edge of the ring which is not traversed by any of the arcs in S.) The size of the partition is defined to be $c + w$. Kumar noticed that coloring the arcs in C with as few colors as possible is equivalent to finding the smallest-sized color partition for C.

LEMMA 7.1. *A family of circular arcs C has a color partition of size k if and only if it can be colored with k colors.*

He also made the following observation.

LEMMA 7.2. *In a parallel routing of I, for every edge e on the ring, there is another edge f such that no arc in the routing goes through both e and f. We say that e and f are complements of each other in the parallel routing.*

Given edge e, Lemma 7.2 asserts that in a parallel-routing optimal solution for unit-DRSP, there exists an edge f such that e and f are complements of each other. Since f is not known in advance, Kumar's algorithm considers each edge of the ring (except, of course edge e) as a candidate for f. The algorithm then tries to find the best solution for unit-DRSP on rings based on the assumption that e and f are complements in an optimal parallel routing. Once all the candidates for f have been considered, the algorithm outputs the best solution.

Let us assume for now that there is a parallel-routing optimal solution for unit-DRSP where edges e and f are complements of each other. Suppose a demand can be routed so that it does not go through both edges e and f. From Lemma 7.2 we know that in the optimal solution this demand must indeed be routed so that it misses edges e and f. Thus, the algorithm considers all the demands in I, and whenever a demand can be routed so that it misses edges e and f, this route is chosen for the demand.

Let I' consist of the routes that have now been routed as well as the remaining demands in I that have not been routed. Let us call the unrouted demands in I' *crossover* demands since, whatever route we choose for these demands, the route crosses either e or f. We shall denote them as $r_1, r_2, \ldots, r_l$. Note that I' consists of arcs and crossover demands.

The algorithm then considers a binary integer program that seeks to find a routing of I' that has the smallest-sized color partition. To approximate the best solution, the continuous relaxation of the program is solved and randomized rounding is performed. We discuss the binary integer program formulation below. We refer the reader to [14], which contains the proof that this formulation outputs the optimal solution to unit-DRSP on rings if e and f are indeed complements in some optimal parallel routing.

Consider the following collection of arcs: $\{a_1, \ldots a_l, a_{l+1}, \ldots, a_{2l}, \ldots, a_m\}$, where a_i and a_{l+i} represent the routes of r_i that go through edges e and f, respectively, for $i = 1, \ldots l$ and $a_{2l+1}, \ldots a_m$ represent the arcs in I'.

Let x_i be the indicator variable that is set to 1 if r_i is routed as a_i for $i = 1, \ldots, l$ and 0 otherwise. Let us adopt the convention that if r_i is routed as a_i, then a_i is assigned the color i too; otherwise, if r_i is routed as a_{l+i}, color i will *not* be used at all. The arc a_{l+i} can, however, be assigned other colors. Thus, x_i also denotes the quantity of color i used.

If the colors $1, \ldots, l$ are not enough to color all the m arcs, then some of them are obliged to remain uncolored; in this case, the arcs are assigned the color 0. (Note that 0 is *not* one of the colors mentioned in the earlier paragraph.) Let $y_{i,j}$ be the

$$\text{Minimize } z = u + \sum_{i=1}^{l} x_i \tag{7.1}$$

subject to

$$y_{i,i} = x_i \qquad \text{for } i = 1, \ldots, l, \tag{7.2}$$

$$\sum_{j=0}^{l} y_{l+i,j} = 1 - x_i \qquad \text{for } i = 1, \ldots, l, \tag{7.3}$$

$$\sum_{j=0}^{l} y_{i,j} = 1 \qquad \text{for } i = 2l+1, \ldots, m, \tag{7.4}$$

$$\sum_{i:g \in a_i} y_{i,j} \leq x_j \qquad \text{for every edge } g,\ j = 1, \ldots, l, \tag{7.5}$$

$$\sum_{i:g \in a_i} y_{i,0} \leq u \qquad \text{for every edge } g, \tag{7.6}$$

$$x_i \in \{0,1\} \qquad \text{for } i = 1, \ldots, l, \tag{7.7}$$

$$y_{i,j} \in \{0,1\} \qquad \text{for } i = 1, \ldots m,\ j = 1, \ldots, l. \tag{7.8}$$

FIG. 7.1. *Kumar's formulation for unit-DRSP with the assumption that edges e and f are complements in an optimal parallel routing.*

variable that is set to 1 if a_i is assigned the color j for $i = 1, \ldots m$, $j = 0, \ldots, l$. The formulation is presented in Figure 7.1.

Constraint (7.2) is based on our convention that, for $1 \leq i \leq l$, if r_i is routed as a_i (i.e., $x_i = 1$), then a_i is assigned the color i. Constraint (7.3) enforces the rule that, for $1 \leq i \leq l$, if r_i is routed as a_i, then no colors are assigned to a_{l+i}. If, on the other hand, r_i is routed as a_{l+i}, then it must be assigned a color. Constraint (7.4) guarantees that each arc a_i, $2l+1 \leq i \leq m$, is assigned a color. Constraint (7.5) guarantees that if color j is available, at most one arc that goes through edge g can be colored j. In the fractional sense, the constraint also requires that no more than x_j of color j should appear on all arcs that go through the edge g. Constraint (7.6) captures the ringload of the collection of uncolored arcs. It is easy to check that for $i = 1, \ldots, l$, the family of arcs colored i has a ringload equal to 1 and that all uncolored arcs collectively have ringload u.

Since solving integer programs exactly is NP-hard, the algorithm considers the continuous relaxation of the formulation above. Let $z^*_{e,f}$ be the objective value of the optimal fractional solution for unit-DRSP on rings where e and f are complements in an optimal parallel routing. Kumar was able to show that there exists at least one such optimal fractional solution with the following property: for i in the set $\{1, \ldots, l\}$ such that $x_i \neq 0$, $y_{l+i,0} = 1 - x_i$. That is, if r_i is routed partly as a_i and partly as a_{l+i} in the fractional solution, then the "fractional arc" a_{l+i} must be uncolored. We shall call this the *uncolorable property.*

Recall that G_I is the graph obtained from viewing the demands of I as chords. We let the vertices of G_I represent the chords and two vertices in G_I were adjacent if and only if their respective chords intersect in the interior of the circle. Let $\omega(G_I) = w$. Without loss of generality, let the chords of the first w' crossover demands and the chords of the first $w - w'$ arcs in I' be the demand chords in I that form a clique of size w in G_I. Since the w chords mutually intersect in the interior of the circle, for $1 \leq j \leq w'$, the route of demand r_j will overlap the routes of $r_1, \ldots, r_{j-1}, r_{j+1}, \ldots, r_{w'}$ and the arcs $a_{2l+1}, \ldots, a_{2l+(w-w')}$. In particular, for $1 \leq j \leq w'$, the arc a_j overlaps with both arcs a_k and a_{l+k}, where $1 \leq k \leq w', k \neq j$; similarly, a_j overlaps with a_{2l+k}, where $1 \leq k \leq w - w'$.

Let the vector $(\vec{x^*}, \vec{y^*})$ be an optimal fractional solution that satisfies the uncolorable property and yields the optimal objective value $z^*_{e,f}$. Let $1 \leq j \leq w'$. In the optimal fractional solution, we claim that none of the arcs in $\{a_{l+1}, \dots, a_{l+w'}, a_{2l+1}, \dots, a_{2l+(w-w')}\}$ can be assigned a positive quantity of color j. If $x^*_j = 0$, the claim is clearly true since color j will not be available at all for coloring any arcs. If $x^*_j > 0$, the arc a_{l+j} cannot be assigned the color j because of the uncolorable property. If some a_{l+k} is assigned a positive quantity of color j, where $1 \leq k \leq w'$ and $k \neq j$, then at least $x^*_j + y^*_{l+k,j}$ amount of color j will appear on all the edges where a_j and a_{l+k} overlap, violating constraint (7.5). For the same reason, none of the arcs in $\{a_{2l+1}, \dots, a_{2l+(w-w')}\}$ can be assigned a positive quantity of the color j. In other words,

$$y^*_{l+k,j} = 0 \qquad \text{for } 1 \leq k \leq w',\ 1 \leq j \leq w', \tag{7.9}$$

$$y^*_{2l+k,j} = 0 \qquad \text{for } 1 \leq k \leq w - w',\ 1 \leq j \leq w'. \tag{7.10}$$

Earlier we said that since the demand chords of the arcs in the set $\{a_{l+1}, \dots, a_{l+w'}\} \cup \{a_{2l+1}, \dots, a_{2l+(w-w')}\}$ mutually intersect in the interior of the circle, the arcs must mutually overlap. In addition, all these arcs do not pass through the edge e. Thus, there must exist an edge of the ring where all these arcs overlap. Let us call this edge g. We have the following inequality when we apply constraint (7.5) on edge g and the colors $j = w' + 1, \dots, l$:

$$\sum_{i=1}^{w'} y^*_{l+i,j} + \sum_{i=2l+1}^{2l+w-w'} y^*_{i,j} \leq x^*_j \qquad \text{for } j = w' + 1, \dots, l. \tag{7.11}$$

We are now ready to show that $\omega(G_I)$ is a lower bound to $z^*_{e,f}$. We note that from constraints (7.2), (7.3), and (7.4) in the formulation,

$$\begin{aligned} y^*_{i,i} + \sum_{j=0}^{l} y^*_{l+i,j} &= 1, \qquad && i = 1, \dots, l, \\ \sum_{j=0}^{l} y^*_{i,j} &= 1, \qquad && i = 2l + 1, \dots, m. \end{aligned}$$

Thus, since $w = w' + (w - w')$, by the last two equations we have

$$\begin{aligned} w &= \sum_{i=1}^{w'} \left(y^*_{i,i} + \sum_{j=0}^{l} y^*_{l+i,j} \right) + \sum_{i=2l+1}^{2l+w-w'} \sum_{j=0}^{l} y^*_{i,j} \\ &= \sum_{i=1}^{w'} y^*_{l+i,0} + \sum_{i=2l+1}^{2l+w-w'} y^*_{i,0} + \sum_{i=1}^{w'} \left(y^*_{i,i} + \sum_{j=w'+1}^{l} y^*_{l+i,j} \right) + \sum_{i=2l+1}^{2l+w-w'} \sum_{j=w'+1}^{l} y^*_{i,j}, \end{aligned}$$

where the second equality follows from (7.9) and (7.10). Hence,

$$\begin{aligned} w &\le u + \sum_{i=1}^{w'} x_i^* + \sum_{j=w'+1}^{l} \left(\sum_{i=1}^{w'} y_{l+i,j}^* + \sum_{i=2l+1}^{2l+w-w'} y_{i,j}^* \right) \\ &\le u + \sum_{i=1}^{w'} x_i^* + \sum_{j=w'+1}^{l} x_j^* \\ &= z_{e,f}^*, \end{aligned}$$

where the first inequality follows from (7.6) and (7.2) and the second inequality from (7.11).

Let E denote the set of edges in the ring. Since $\omega(G_I) \le z_{e,f}^*$ for each $f \in E$, we have the following theorem.

THEOREM 7.3. *Let I be a set of demands on the ring and $z_{frac}^* = \min_{f\in E} z_{e,f}^*$, where $z_{e,f}^*$ is the optimal fractional solution to the binary integer program for unit-DRSP on rings which assumes that edges e and f are complements in an optimal parallel routing of unit-DRSP on rings. Let $\omega(G_I)$ denote the size of the largest clique in G_I. Then $\omega(G_I) \le z_{frac}^*$.*

7.3. Bounding the number of slots used in HRA. Our goal in this section is to show that there exists an M so that when $z_0(n) = M \ln n$,

$$1.5 + 1/(2e) + f(n, z_0(n)) \le 2 - 1/\lceil z_0(n)/2 \rceil. \tag{7.12}$$

That is, w.h.p., the approximation factor of Kumar's algorithm is better than that of the Unit-DRSP Algorithm.

First, let us assume that $n \ge 2$ since a ring should have at least two nodes. Also assume that $z_0(n)^3 \ln z_0(n) \ge 1$ so that $8/z_0^2(n) \le 8\sqrt{\ln z_0(n)/z_0(n)}$. Thus,

$$\begin{aligned} 1.5 + \frac{1}{2e} + f(n, z_0(n)) &= 1.5 + \frac{1}{2e} + 2\sqrt{2(\ln n + 2\ln z_0(n))/z_0(n)} \\ &\quad + \sqrt{2\ln z_0(n)/z_0(n)} + 8/(z_0(n))^2 \\ &\le 1.69 + (12 + \sqrt{2})\sqrt{\frac{\ln n + \ln z_0(n)}{z_0(n)}}. \\ &= F_1(n, z_0(n)). \end{aligned}$$

Let $F_2(z_0(n)) = 2 - 2/z_0(n) \le 2 - 1/\lceil z_0(n)/2 \rceil$. We note that if $n \ge 2$, $z_0(n)^3 \ln z_0(n) \ge 1$, and $F_1(n, z_0(n)) \le F_2(z_0(n))$, then (7.12) is true. Let us consider a necessary and sufficient condition for $F_1(n, z_0(n)) \le F_2(z_0(n))$:

$$\begin{aligned} 1.69 + (12 + \sqrt{2})\sqrt{\frac{\ln n + \ln z_0(n)}{z_0(n)}} &\le 2 - \frac{2}{z_0(n)} \\ \Leftrightarrow \sqrt{\frac{\ln n + \ln z_0(n)}{z_0(n)}} &\le \frac{1}{12 + \sqrt{2}}\left(0.31 - \frac{2}{z_0(n)}\right) \\ \Leftrightarrow \ln n &\le \frac{z_0(n)}{(12 + \sqrt{2})^2}\left(0.31 - \frac{2}{z_0(n)}\right)^2 - \ln z_0(n). \end{aligned}$$

Thus, we have established the following proposition.

Proposition 7.4. *Let* $K_0 = (\frac{0.31}{12+\sqrt{2}})^2$, $K_1 = \frac{4(0.31)}{(12+\sqrt{2})^2}$, *and* $K_2 = (\frac{2}{12+\sqrt{2}})^2$. *Assume* $n \geq 2$. *If*

(a) $z_0(n)^3 \ln z_0(n) \geq 1$ *and*

(b) $\ln n \leq K_0 z_0(n) - K_1 + K_2/z_0(n) - \ln z_0(n)$,

then $1.5 + 1/(2e) + f(n, z_0(n)) \leq 2 - 1/\lceil z_0(n)/2 \rceil$.

Earlier, we set $z_0(n) = M \ln n$. Let us now define M so that conditions (a) and (b) of Proposition 7.4 are true. To find M, we note that as $x \to \infty$, the fraction $\frac{x}{\ln x} \to \infty$. Since K_0 and K_1 are constants, there must exist an M_0 such that $\frac{K_0 M_0 - 2}{\ln M_0 - K_1} > \frac{1}{\ln 2}$. Set $M = \max\{M_0, e/\ln 2\}$. Hence, $M \ln n \geq e$ so $z_0(n)^3 \ln z_0(n) \geq 1$. Furthermore,

$$\begin{aligned}
K_0 M \ln n &- K_1 + K_2/(M \ln n) - \ln(M \ln n) \\
&= K_0 M \ln n - K_1 + K_2/(M \ln n) - \ln M - \ln \ln n \\
&\geq \ln n + \ln n - \ln \ln n + (K_0 M - 2) \ln n - \ln M - K_1 \\
&\geq \ln n + \ln n - \ln \ln n + (K_0 M - 2) \ln 2 - \ln M - K_1 \\
&\geq \ln n,
\end{aligned}$$

where the first and second inequalities are true because $n \geq 2$, so $K_2/(M \ln n) > 0$ and $\ln n \geq \ln 2$ and the third inequality follows from the choice of M and the fact that $\ln n > \ln \ln n$ when $n \geq 2$. From Proposition 7.4, we have proved that our choice for M makes the inequality $1.5 + 1/(2e) + f(n, z_0(n)) \leq 2 - 1/\lceil z_0(n)/2 \rceil$ true. So when $z^*_{frac} = z_0(n) = \max\{M_0, e/\ln 2\} \ln n$, w.h.p., the approximation factor of Kumar's algorithm is not more than that of Unit-DRSP.

Acknowledgments. I am very grateful for Lenore Cowen's advice and encouragement throughout the course of this work. I would also like to thank Leslie Hall and Tami Carpenter for many helpful discussions on topics related to the paper, Dany Kleitman and Tom Leighton for introducing me to DRSP, Vijay Kumar for preprints of his two papers, and the anonymous referees whose recommendations greatly improved the presentation of this paper.

REFERENCES

[1] B. Beauquier, J.-C Bermond, L. Gargano, P. Hell, S. Perennes, and U. Vaccaro, *Graph problems arising from wavelength-routing in all-optical networks*, presented at the 2nd IEEE Workshop on Optics and Comput. Science, Geneva, Switzerland, 1997.

[2] J.-C Bermond, L. Gargano, S. Perennes, A. A. Rescigno, and U. Vaccaro, *Efficient Collective Communication in Optical Networks*, Lecture Notes in Comput. Sci. 1099, Springer-Verlag, New York, 1996, pp. 574–585.

[3] B. Bhattacharya and D. Kaller, *An $O(m + n \log n)$ algorithm for the maximum-clique problem in circular-arc graphs*, J. Algorithms, 25 (1997), pp. 336–358.

[4] T. Carpenter, *personal communication*, Telcordia Technologies, Morristown, NJ, 1999.

[5] T. Carpenter, S. Cosares, and I. Saniee, *Demand Routing and Slotting on Ring Networks*, Technical report 97-02, Bellcore, Morristown, NJ, 1997.

[6] T. Erlebach and K. Jansen, *Call scheduling in trees, rings and meshes*, in Proceedings of the 30th Hawaii International Conference on System Sciences Vol. 1, IEEE Computer Society, Los Alamitos, CA, 1997, pp. 221–222.

[7] A. Frank, *Edge-disjoint paths in planar graphs*, J. Combin. Theory Ser. B, 39 (1985), pp. 164–178.

[8] M. Garey, D. Johnson, G. Miller, and C. Papadimitriou, *The complexity of coloring circular arcs and chords*, SIAM J. Algebraic Discrete Methods, 1 (1980), pp. 216–227.

[9] F. Gavril, *Algorithms for a maximum clique and a maximum independent set of circle graphs*, Networks, 3 (1973), pp. 261–273.

[10] M. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[11] W. L. Hsu and W. K. Shih, *An approximation algorithm for coloring circular arc graphs*, presented at the SIAM Conference on Discrete Mathematics, Atlanta, GA, 1990.

[12] I. A. Karapetian, *On coloring arc graphs*, Dokladi (Reports) of the Academy of Science of the Armenian Soviet Socialist Republic, 70 (1980), pp. 306–311.

[13] V. Kumar, *An approximation algorithm for circular arc coloring*, Algorithmica, 31 (2001), pp. 406–417.

[14] V. Kumar, *Approximating circular arc coloring and bandwidth allocation in all-optical ring networks*, in Approximation Algorithms for Combinatorial Optimization, Lecture Notes in Comput. Sci. 1444, K. Jansen and J. Rolim, eds., Springer-Verlag, Berlin, 1998, pp. 147–158.

[15] V. Kumar and E. J. Schwabe, *Improved access to optical bandwidth in trees*, in Proceedings of the 8th Annual ACM–SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 1997, pp. 437–444.

[16] A. Litman and A. L. Rosenberg, *Balancing Communication in Ring-Structured Networks*, Technical report 93-80, University of Massachusetts, Amherst, MA, 1993.

[17] M. Mihail, C. Kaklamanis, and S. Rao, *Efficient access to optical bandwidth*, in Proceedings of the 36th IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, Los Alamitos, CA, 1995, pp. 548–557.

[18] P. Raghavan and E. Upfal, *Efficient routing in all-optical networks*, in Proceedings of the 26th ACM Symposium on Theory of Computing, ACM, New York, 1994, pp. 134–143.

[19] A. Schrijver, P. Seymour, and P. Winkler, *The ringloading problem*, SIAM J. Discrete Math., 11 (1998), pp. 1–14.

[20] A. Tucker, *Coloring a family of circular arcs*, SIAM J. Appl. Math., 29 (1975), pp. 493–502.

[21] G. Wilfong and P. Winkler, *Ring routing and wavelength translation*, in Proceedings of the 9th Annual ACM–SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 1998, pp. 333–341.

SIAM J. Discrete Math.
Vol. 17, No. 3, pp. 403–425

GRAPH IMPERFECTION WITH A CO-SITE CONSTRAINT*

STEFANIE GERKE[†] AND COLIN MCDIARMID[‡]

Abstract. We are interested in a version of graph coloring where there is a "co-site" constraint value k. Given a graph G with a nonnegative integral demand x_v at each node v, we must assign x_v positive integers (colors) to each node v such that the same integer is never assigned to adjacent nodes, and two distinct integers assigned to a single node differ by at least k. The aim is to minimize the span, that is, the largest integer assigned to a node. This problem is motivated by radio channel assignment where one has to assign frequencies to transmitters so as to avoid interference. We compare the span with a clique-based lower bound when some of the demands are large. We introduce the relevant graph invariant, the k-imperfection ratio, give equivalent definitions, and investigate some of its properties. The k-imperfection ratio is always at least 1: we call a graph k-perfect when it equals 1. Then 1-perfect is the same as perfect, and we see that for many classes of perfect graphs, each graph in the class is k-perfect for all k. These classes include comparability graphs, co-comparability graphs, and line-graphs of bipartite graphs.

Key words. imperfection ratio, generalized graph coloring, perfect graphs, channel assignment

AMS subject classifications. 05C15, 05C17, 05C80, 05C90, 90C27

DOI. 10.1137/S0895480102402514

1. Introduction.

We are interested in a problem motivated by radio channel assignment in cellular networks, where one has to assign sets of frequencies or channels to transmitters so as to satisfy the local demand for channels at every transmitter, to avoid unacceptable interference, and to use the minimum amount of the spectrum; see, for example, [13], [14], or [17]. We assume that the interference is acceptable if any two channels assigned to a pair of potentially interfering transmitters are different and the distance (in the spectrum) between two distinct channels assigned to the same transmitter is at least k, where the positive integer k is a given constant which is called the *co-site constraint value.* Typically k will be a small positive integer. We are particularly interested in this problem when the demand for channels at some of the transmitters is large. This is not only because this case is important in practical situations, but even more since it leads to significant simplifications that reveal interesting structure.

If we represent colors by positive integers $1, 2, \ldots$, then this problem translates to coloring the nodes of a weighted graph $G = (V, E)$ with nonnegative integral weight vector $\mathbf{x} = (x_v : v \in V)$ in such a way that x_v colors are assigned to each node v, two colors assigned to adjacent nodes are different, and two distinct colors assigned to the same node differ by at least the co-site constraint value k. Such a coloring is called *k-feasible* for G and $\mathbf{x}$. The objective is to minimize the largest number used. We define $\text{span}_k(G, \mathbf{x})$ to be the minimum value of the largest number used, over all k-feasible assignments for G and $\mathbf{x}$. Observe that $\text{span}_1(G, \mathbf{1})$ equals the chromatic number $\chi(G)$ of the graph G.

*Received by the editors February 12, 2002; accepted for publication (in revised form) June 25, 2003; published electronically January 6, 2004.
http://www.siam.org/journals/sidma/17-3/40251.html

[†]TU München, Institut für Informatik, Bolzmannstraße 3, D-85748 Garching (gerke@in.tum.de).
The research of the author was partially supported by the EPSRC under grant 970004215 and the UK Radiocommunication Agency.

[‡]University of Oxford, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG (cmcd@stats.ox.ac.uk).

We want to compare $\mathrm{span}_k(G, \mathbf{x})$ with a clique-based lower bound, when some of the demands are large. To do this we set

$$\omega_k(G, \mathbf{x}) = \max\left\{\mathrm{span}_k(K, \mathbf{x}) : K \text{ is a clique in } G\right\},$$

where we abuse notation and use $\mathbf{x}$ also for its restriction to subgraphs of G. It is known [9] that for a clique K, there is a simple formula for $\mathrm{span}_k(K, \mathbf{x})$; see (2.1) below. Observe that $\omega_k(G, \mathbf{x})$ is always at least $\omega(G, \mathbf{x})$, where $\omega(G, \mathbf{x})$ is the maximum of $\sum_{v \in V(K)} x_v$ over all cliques K of G.

Given a weight vector $\mathbf{x}$ of a graph G, that is, a nonnegative vector indexed by the nodes of G, we let $x_{\max}$ denote the maximum value of x_v over all the nodes v. We set

$$s_k^j(G) = \max\left\{\frac{\mathrm{span}_k(G, \mathbf{x})}{\omega_k(G, \mathbf{x})} : x_{\max} = j\right\}, \tag{1.1}$$

where the maximum is over all integral weight vectors with $x_{\max} = j$. Observe that $s_k^j(G) \geq 1$ by definition, and $s_k^j(G) = 1$ if G is a complete graph. Consider the case $k = 1$: it is known [7] that $\lim_{j\to\infty} s_1^j(G)$ exists and is the imperfection ratio, which we discuss below. We will see that the corresponding result holds for each positive integral k, namely $s_k^j(G)$ tends to a limit as $j \to \infty$. This limit is the "k-imperfection ratio" and is the subject of this paper. In order to give a convenient definition of it, we first introduce the fractional k-clique-bound and the fractional k-span.

For a fixed positive integer k, the *fractional k-clique-bound* $\omega_k^f(G, \mathbf{x})$ of a graph G with weight vector $\mathbf{x}$ is

$$\omega_k^f(G, \mathbf{x}) = \max\left\{k x_{\max}, \omega(G, \mathbf{x})\right\},$$

and the *fractional k-span* $\mathrm{span}_k^f(G, \mathbf{x})$ is the value of the following linear program (LP) which has a variable y_S for each induced k-colorable subgraph S of G: $\min k \sum_S y_S$ subject to $\sum_{S \ni v} y_S \geq x_v$ for each node v, and $y_S \geq 0$ for each k-colorable induced subgraph S of G. Observe that for any graph G, $\mathrm{span}_1^f(G, \mathbf{x})$ is the weighted fractional chromatic number $\chi_f(G, \mathbf{x})$, and, in particular, $\mathrm{span}_1^f(G, \mathbf{1})$ is the fractional chromatic number $\chi_f(G)$. It is easy to check that

$$\mathrm{span}_k^f(G, \mathbf{x}) \geq \omega_k^f(G, \mathbf{x}), \tag{1.2}$$

and we do so toward the end of this introductory section. Observe that one can easily extend the definitions of $\omega_k^f(G, \mathbf{x})$ and $\mathrm{span}_k^f(G, \mathbf{x})$ to rational or real weight vectors, and we will use them in this way later, whereas we will use $\omega_k(G, \mathbf{x})$ and $\mathrm{span}_k(G, \mathbf{x})$ for integral weight vectors only.

The *k-imperfection ratio* $\mathrm{imp}_k(G)$ of a graph G is defined by setting

$$\mathrm{imp}_k(G) = \sup_{\mathbf{x}} \frac{\mathrm{span}_k^f(G, \mathbf{x})}{\omega_k^f(G, \mathbf{x})}, \tag{1.3}$$

where the supremum is over all nonzero integral weight vectors $\mathbf{x}$. It turns out that there always exists such a vector $\mathbf{x}$ with $\mathrm{imp}_k(G) = \mathrm{span}_k^f(G, \mathbf{x})/\omega_k^f(G, \mathbf{x})$ (see (2.7) below), and thus the supremum in the definition (1.3) may be replaced by the maximum. Observe that if H is an induced subgraph of G, then $\mathrm{imp}_k(H) \leq \mathrm{imp}_k(G)$. By (1.2) we have

$$\mathrm{imp}_k(G) \geq 1. \tag{1.4}$$

We say that a graph G with $\mathrm{imp}_k(G) = 1$ is *k-perfect.* Observe that if $\chi(G) \leq k$, then trivially G is k-perfect, since then $\mathrm{span}_k^f(G, \mathbf{x}) = k x_{\max}$ (take $y_V = x_{\max}$).

The 1-imperfection-ratio has been studied in [7], [8]. It is called the *imperfection ratio* and is denoted by $\mathrm{imp}(G)$. Its name was motivated by the fact that $\mathrm{imp}(G) \geq 1$ for all graphs G and that $\mathrm{imp}(G) = 1$ if and only if G is perfect. We shall see that not every perfect graph is k-perfect when $k \geq 2$, but, for example, this does hold for comparability graphs and some other classes of graphs, and there are many interesting properties of the imperfection ratio which have their equivalents in the more general case.

The plan of the paper is as follows. After giving three introductory results at the end of this section, we see in section 2 that for any graph G and any positive integer k, the quantity s_k^j defined in (1.1) above satisfies

$$s_k^j(G) \to \mathrm{imp}_k(G) \text{ as } j \to \infty, \tag{1.5}$$

and we present equivalent alternative polyhedral definitions of $\mathrm{imp}_k(G)$.

In section 3 we find upper and lower bounds on the k-imperfection ratio, including the result that $\mathrm{imp}_k(G)/\mathrm{imp}(G) \leq \frac{1}{1+1/e} \sim 1.6$. These bounds yield some extremal results. We also see, for example, that the Petersen graph P satisfies $\mathrm{imp}_2(P) = 10/7$.

In section 4 we see that the class of 2-perfect graphs is a proper subclass of the class of perfect graphs. In contrast, it is easy to find nonperfect graphs which are k-perfect for each $k \geq 3$, for example, the odd cycles C_n on $n \geq 5$ nodes (the *odd holes*). We then consider some classes of perfect graphs, where each graph G in the class is k-perfect for each positive integer k. We call such a graph G *all-perfect*. We already know that this holds for bipartite graphs (since $\chi(G) \leq k$ for each $k \geq 2$), and we shall see shortly that it is true also for complete graphs. In section 4 we shall see that it is also true for comparability graphs, co-comparability graphs, and line-graphs of bipartite graphs.

In section 5 we see that, in contrast to the nice behavior for perfect graphs, for each $k \geq 2$ there are many nonisomorphic node-minimal non-k-perfect graphs: indeed, the number on at most n nodes grows at least exponentially with n. We see that an odd hole on n nodes is node-minimal non-2-perfect and that its complement (an *odd antihole*) is node-minimal non-k-perfect for all $k \leq (n-1)/2$. We also determine imp_k for all odd holes and antiholes.

In section 6 we consider disk graphs, which crop up naturally in models for radio channel assignment, and give bounds for their k-imperfection ratio.

In section 7 we see that for the random graph $G_{n,\frac{1}{2}}$, the k-imperfection ratio is about $n/(4\log_2^2 n)$ (which is independent of k), and also we obtain corresponding results for sparse random graphs and random regular graphs (which do depend on k).

Let us finish this section by giving three simple introductory results, as mentioned above. The first task is to prove (1.2). Let (y_S) be a feasible solution to the LP defining $\mathrm{span}_k^f(G, \mathbf{x})$. If x_v is $x_{\max}$, then

$$k \sum_S y_S \geq k \sum_{S: v \in S} y_S \geq k x_v = k x_{\max}.$$

Also, if the set K of nodes forms a complete subgraph of G with $\omega(G, \mathbf{x}) = \sum_{v \in K} x_v$, then, since $|K \cap S| \leq k$ for each k-colorable subset S, we have

$$k \sum_S y_S \geq \sum_S \sum_{v \in K \cap S} y_S = \sum_{v \in K} \sum_{S: v \in S} y_S \geq \sum_{v \in K} x_v = \omega(G, \mathbf{x}).$$

Hence $k \sum_S y_S \geq \omega_k^f(G, \mathbf{x})$, which yields (1.2).

The second of our three introductory results shows that (not unexpectedly) we may restrict our attention to connected graphs.

Proposition 1.1. *For any positive integer k and any graph G, if G consists of the disjoint union of graphs $G_1, \ldots, G_t$, then*

$$\mathrm{imp}_k(G) = \max\{\mathrm{imp}_k(G_1), \ldots, \mathrm{imp}_k(G_t)\}.$$

Proof. Directly from the definitions, for any weight vector $\mathbf{x}$

$$\begin{aligned}\mathrm{span}_k^f(G,\mathbf{x}) &= \max_i\{\mathrm{span}_k^f(G_i,\mathbf{x})\}\\ &\le \max_i\{\mathrm{imp}_k(G_i)\ \omega_k^f(G_i,\mathbf{x})\}\\ &\le (\max_i \mathrm{imp}_k(G_i))\ \omega_k^f(G,\mathbf{x}),\end{aligned}$$

and so $\mathrm{imp}_k(G) \le \max_i \mathrm{imp}_k(G_i)$. The lower bound follows immediately from the earlier remark that the k-imperfection ratio of an induced subgraph of G is always at most the k-imperfection ratio of G. □

Next we meet a connection with scheduling theory. Recall that we say that a graph is all-perfect if it is k-perfect for each positive integer k.

Proposition 1.2. *Each complete graph K is all-perfect.*

Proof. This result will follow directly from the fact we noted above (see (2.6) below) that $1 = s_k^j(K) \to \mathrm{imp}_k(K)$ as $j \to \infty$, but it is interesting to note that it is a disguised form of a standard basic result in scheduling theory. Suppose that we have k identical machines in parallel, a collection V of jobs v with processing time x_v, and pre-emptions are allowed (we need at most 1 per job). It is well known [19] and not hard to see that the makespan m (the minimum completion time) is given by

$$m = \max\left\{x_{\max}, \left(\sum_v x_v\right)/k\right\} = \omega_k^f(K,\mathbf{x})/k.$$

Given a schedule with makespan m, for each set $S \subseteq V$ let y_S be the total time that S is the set of jobs being processed. Then $\sum_{S:v\in S} y_S = x_v$ for each $v \in V$, and $\sum_S y_S = m$. □

2. Equivalent descriptions. In this section we introduce equivalent polyhedral descriptions for $\mathrm{imp}_k(G)$; see Theorem 2.3. We also show that for any graph G there is an integral weight vector $\mathbf{x}$ with $\mathrm{imp}_k(G) = \mathrm{span}_k^f(G,\mathbf{x})/\omega_k^f(G,\mathbf{x})$ and each coordinate at most $2^{-n}(n+1)^{(n+1)/2}$ (where $n = |V(G)|$). It was shown in [7] that we may need coordinates as large as $2^{(n-5)/4}$ if $k = 1$.

To prove Theorem 2.3 we need one preliminary lemma, some more notation, and a result of [9], which says that for any clique K and any integral weight vector $\mathbf{x}$,

$$\mathrm{span}_k(K,\mathbf{x}) = \max\left\{(x_{\max}-1)k + |\{v \in V(K) : x_v = x_{\max}\}|, \sum_{v\in V(K)} x_v\right\}. \tag{2.1}$$

By this result,

$$\omega_k^f(G,\mathbf{x}) - k + 1 \le \omega_k(G,\mathbf{x}) \le \omega_k^f(G,\mathbf{x}), \tag{2.2}$$

and since $\omega_k^f(G,a\mathbf{x}) = a\,\omega_k^f(G,\mathbf{x})$,

$$\omega_k^f(G,\mathbf{x}) = \lim_{a\to\infty}\frac{\omega_k(G,a\mathbf{x})}{a}.$$

The latter equality partially motivated the notation $\omega_k^f(G, \mathbf{x})$ since many fractional versions of graph parameters can be defined in this way [21]—see also Corollary 2.2 below.

LEMMA 2.1. *For any graph G on n nodes with integral weight vector $\mathbf{x}$,*

$$\mathrm{span}_k^f(G, \mathbf{x}) - k \le \mathrm{span}_k(G, \mathbf{x}) \le \mathrm{span}_k^f(G, \mathbf{x}) + 2nk.$$

Proof. In any k-feasible assignment each node belongs to at most one of any k consecutive color classes, and the graph induced by the nodes of k consecutive color classes is k-colorable. Hence each node v can be covered x_v times by at most $\lceil \mathrm{span}_k(G, \mathbf{x})/k \rceil$ k-colorable graphs, which yields

$$\frac{\mathrm{span}_k^f(G, \mathbf{x})}{k} \le \left\lceil \frac{\mathrm{span}_k(G, \mathbf{x})}{k} \right\rceil \le \frac{\mathrm{span}_k(G, \mathbf{x}) + k - 1}{k},$$

and so

$$\mathrm{span}_k^f(G, \mathbf{x}) - k + 1 \le \mathrm{span}_k(G, \mathbf{x}).$$

To prove that $\mathrm{span}_k(G, \mathbf{x}) \le \mathrm{span}_k^f(G, \mathbf{x}) + 2nk$, consider an optimal basic feasible solution $\mathbf{y}$ of the LP determining $\mathrm{span}_k^f(G, \mathbf{x})$. Since $\mathbf{y}$ is a basic feasible solution, at most n values y_S are nonzero. Hence by rounding up $\mathbf{y}$ one obtains an integral feasible solution $\mathbf{z}$ with value less than $\mathrm{span}_k^f(G, \mathbf{x}) + nk$. Now we can color a k-colorable subgraph S of G in a k-feasible way z_S times using $z_S k$ consecutive colors. To put these colorings together for a k-feasible assignment one can introduce gaps of size $k - 1$ to ensure that two distinct colors assigned to a node are at least k apart. Hence

$$\mathrm{span}_k(G, \mathbf{x}) < \mathrm{span}_k^f(G, \mathbf{x}) + nk + (n-1)(k-1) \le \mathrm{span}_k^f(G, \mathbf{x}) + 2nk$$

as claimed. □

Since $\mathrm{span}_k^f(G, a\mathbf{x}) = a\,\mathrm{span}_k^f(G, \mathbf{x})$, Lemma 2.1 yields as a corollary the following result, which motivated the choice of the name "fractional k-span."

PROPOSITION 2.2. $\mathrm{span}_k^f(G, \mathbf{x}) = \lim_{a \to \infty} \mathrm{span}_k(G, a\mathbf{x})/a$.

We denote the set of all real weight vectors $\mathbf{x}$ with $\omega_k^f(G, \mathbf{x}) \le 1$ by $QSTAB_k(G)$, or equivalently

$$QSTAB_k(G) = QSTAB(G) \cap [0, 1/k]^n,$$

where $QSTAB(G) = QSTAB_1(G)$ is the *fractional node-packing polytope*; see [12] for further discussion. The convex hull of the incidence vectors of the k-colorable induced subgraphs of G scaled by $1/k$ is denoted by $STAB_k(G)$. Thus $STAB_1(G)$ is the familiar stable set polytope; again see [12] for further discussion. Note that

$$\mathrm{span}_k^f(G, \mathbf{x}) \le t \text{ if and only if } \mathbf{x} \in t\,STAB_k(G). \tag{2.3}$$

Here $t\,P$ denotes the scaled set $\{t\mathbf{x} : \mathbf{x} \in P\}$. We are now able to state and prove the main theorem of this section.

THEOREM 2.3. *For any graph G,*

$$\mathrm{imp}_k(G) = \min\{t : QSTAB_k(G) \subseteq t\,STAB_k(G)\} \tag{2.4}$$

$$= \max\{\mathrm{span}_k^f(G, \mathbf{x}) : \mathbf{x} \textit{ is a vertex of } QSTAB_k(G)\} \tag{2.5}$$

$$= \lim_{j \to \infty} s_k^j(G). \tag{2.6}$$

In addition, there exists an integral weight vector $\mathbf{x}$ *with*

$$\operatorname{imp}_k(G) = \frac{\operatorname{span}_k^f(G, \mathbf{x})}{\omega_k^f(G, \mathbf{x})}, \tag{2.7}$$

and if G has n nodes, then there is such a vector $\mathbf{x}$ *with each coordinate at most* $2^{-n}(n+1)^{(n+1)/2}$.

Proof. Let $s(G)$ denote the right-hand side of (2.4). Observe that

$$s(G) = \min\{t : \mathbf{x} \in t\, STAB_k(G) \text{ for all vertices } \mathbf{x} \in QSTAB_k(G)\},$$

which equals (2.5) because of (2.3). Thus $s(G)$ is rational, and $QSTAB_k(G) \subseteq s(G)\, STAB_k(G)$. Consider a weight vector $\mathbf{x}$, with $\omega_k^f(G, \mathbf{x}) = l$, which implies that $\mathbf{x} \in l\, QSTAB_k(G)$, so $\mathbf{x} \in ls(G)\, STAB_k(G)$, and hence $\operatorname{span}_k^f(G, \mathbf{x}) \le ls(G)$. Thus $\operatorname{span}_k^f(G, \mathbf{x})/\omega_k^f(G, \mathbf{x}) \le s(G)$, and it follows that $\operatorname{imp}_k(G) \le s(G)$.

Now we show that $\operatorname{imp}_k(G) \ge s(G)$. Let $\mathbf{x}$ be a vertex of $QSTAB_k(G)$ such that $s(G) = \operatorname{span}_k^f(G, \mathbf{x})$. Since $\mathbf{x}$ is rational, we may choose a positive integer N such that the vector $\tilde{\mathbf{x}} = N\mathbf{x}$ is integral. Then $\operatorname{span}_k^f(G, \tilde{\mathbf{x}}) = Ns(G)$, and $\omega_k^f(G, \tilde{\mathbf{x}}) \le N$. Hence $\operatorname{imp}_k(G) \ge \operatorname{span}_k^f(G, \tilde{\mathbf{x}})/\omega_k^f(G, \tilde{\mathbf{x}}) \ge s(G)$. Thus $s(G) = \operatorname{imp}_k(G)$, and further the supremum of the ratios $\operatorname{span}_k^f(G, \mathbf{x})/\omega_k^f(G, \mathbf{x})$ over all weight vectors $\mathbf{x}$ as in the definition of $\operatorname{imp}_k(G)$ is attained at $\tilde{\mathbf{x}}$ (and thus at all integer multiples of $\tilde{\mathbf{x}}$).

Next we prove (2.6). Let $\tilde{\mathbf{x}}$ be an integral weight vector as above such that $\operatorname{span}_k^f(G, \tilde{\mathbf{x}})/\omega_k^f(G, \tilde{\mathbf{x}}) = \operatorname{imp}_k(G)$. Let u be a node of maximal demand, and let $\tilde{l} = \tilde{x}_{\max} = \tilde{x}_u$. For any $l \ge \tilde{l}$, write $l = q\tilde{l} + r$ with $0 \le r < \tilde{l}$, and define $\mathbf{y}_u^l = l = q\tilde{x}_u + r$ and $\mathbf{y}_v^l = q\tilde{x}_v$ for all $v \ne u$. We have $y_{\max}^l = l$ and thus

$$\begin{aligned}
s_k^l(G) &\ge \frac{\operatorname{span}_k(G, \mathbf{y}^l)}{\omega_k(G, \mathbf{y}^l)} \ge \frac{\operatorname{span}_k(G, q\tilde{\mathbf{x}})}{\omega_k(G, q\tilde{\mathbf{x}}) + rk} \ge \frac{\operatorname{span}_k^f(G, q\tilde{\mathbf{x}}) - k}{\omega_k^f(G, q\tilde{\mathbf{x}}) + rk + 2nk} \\
&\ge \frac{\operatorname{span}_k^f(G, q\tilde{\mathbf{x}})}{\omega_k^f(G, q\tilde{\mathbf{x}})} \frac{\omega_k^f(G, q\tilde{\mathbf{x}})}{\omega_k^f(G, q\tilde{\mathbf{x}}) + \tilde{l}k + 2nk} - \frac{k}{\omega_k^f(G, q\tilde{\mathbf{x}})} \\
&\ge \operatorname{imp}_k(G) \frac{k(l - \tilde{l})}{kl + 2nk} - \frac{k}{k(l - \tilde{l})}.
\end{aligned}$$

Now, let $\mathbf{x}^l$ be a weight vector such that $x_{\max}^l = l$ and $s_k^l(G) = \operatorname{span}_k(G, \mathbf{x}^l)/\omega_k(G, \mathbf{x}^l)$. We obtain

$$\begin{aligned}
s_k^l &= \frac{\operatorname{span}_k(G, \mathbf{x}^l)}{\omega_k(G, \mathbf{x}^l)} \le \frac{\operatorname{span}_k^f(G, \mathbf{x}^l) + 2nk}{\omega_k^f(G, \mathbf{x}^l) - k} \\
&\le \frac{\operatorname{span}_k^f(G, \mathbf{x}^l)}{\omega_k^f(G, \mathbf{x}^l)} \frac{\omega_k^f(G, \mathbf{x}^l)}{\omega_k^f(G, \mathbf{x}^l) - k} + \frac{2nk}{\omega_k^f(G, \mathbf{x}^l) - k} \\
&\le \operatorname{imp}(G) \frac{kl}{kl - k} + \frac{2nk}{kl - k},
\end{aligned}$$

and the result follows.

It remains to show that there is weight vector $\mathbf{x}$ as in (2.7) with "small" coordinates. Any vertex $\mathbf{y}$ of $QSTAB_k(G)$ is the unique solution of $A\mathbf{z} = \mathbf{b}$ for some $n \times n$ matrix A with $0, 1$ entries and some vector $\mathbf{b}$ the entries of which equal $0, 1$

or $1/k$. Therefore and because $y_i \leq 1/k$, Cramer's rule implies that $\mathbf{y}$ has entries of the form $a_i/(k\det(A))$ for integers $0 \leq a_i \leq \det(A)$, $i = 1, \ldots, n$. But since A is a 0, 1-matrix, $\det(A) \leq 2^{-n}(n+1)^{(n+1)/2}$ [1]. Considering a vertex $\mathbf{y}$ of $QSTAB_k(G)$ with $\mathrm{imp}_k(G) = \mathrm{span}_k^f(G, \mathbf{y})$ and setting $\mathbf{x} = k\det(A)\mathbf{y}$ yields the result. □

Let us note one more equivalent definition of the k-imperfection ratio, which follows easily from the work above. We could define $\mathrm{imp}_k(G)$ as the least a such that (2.8) below holds for some choice of b.

PROPOSITION 2.4. *Consider a graph G and a positive integer k. Let A be the set of values a such that, for some b,*

$$\mathrm{span}_k(G, \mathbf{x}) \leq a\,\omega_k(G, \mathbf{x}) + b \quad \textit{for each integral weight vector } \mathbf{x}. \tag{2.8}$$

Then $a \in A$ if and only if $a \geq \mathrm{imp}_k(G)$.

Proof. If G has n nodes, by Lemma 2.1 and (2.2)

$$\begin{aligned}\mathrm{span}_k(G, \mathbf{x}) &\leq \mathrm{span}_k^f(G, \mathbf{x}) + 2nk\\ &\leq \mathrm{imp}_k(G)\,\omega_k^f(G, \mathbf{x}) + 2nk\\ &\leq \mathrm{imp}_k(G)\,\omega_k(G, \mathbf{x}) + \mathrm{imp}_k(G)(k-1) + 2nk.\end{aligned}$$

Thus there is a constant b such that (2.8) holds. Conversely, suppose that a and b are such that (2.8) holds. Then $s_k^j(G) \leq a + b/j$, and so by (2.6) it follows that $\mathrm{imp}_k(G) \leq a$. □

3. Bounds.

In this section, we first give bounds on $\mathrm{imp}_k(G)$ in terms of the $\chi(G)$, $\chi_f(G)$, $\omega(G)$, and so on. From these bounds we make various deductions, including determining the value of $\mathrm{imp}_2(P)$ for the Petersen graph P. Next we give an upper bound on $\mathrm{imp}_k(G)$ in terms of $\mathrm{imp}(G)$. These results, together with results from [8], yield various extremal results.

LEMMA 3.1. *For any positive integer k and any graph G,*

(a) $\mathrm{imp}_k(G) \geq \min\{\chi_f(G)/k, \chi_f(G)/\omega(G)\}$,

(b) $\mathrm{imp}_k(G) \leq \mathrm{span}_k^f(G, \mathbf{1})/k \leq \max\{1, \chi(G)/k\}$.

Proof. We have

$$\mathrm{imp}_k(G) \geq \frac{\mathrm{span}_k^f(G, \mathbf{1})}{\omega_k^f(G, \mathbf{1})} \geq \frac{\chi_f(G)}{\max\{\omega(G), k\}} = \min\left\{\frac{\chi_f(G)}{k}, \frac{\chi_f(G)}{\omega(G)}\right\}$$

as required for (a).

For every $\mathbf{x}$ in $QSTAB_k(G)$, we have $\mathbf{x} \leq (1/k)\mathbf{1}$ and therefore

$$\mathrm{span}_k^f(G, \mathbf{x}) \leq \mathrm{span}_k^f(G, (1/k)\mathbf{1}) = \frac{\mathrm{span}_k^f(G, \mathbf{1})}{k}.$$

The first inequality of (b) now follows by (2.5). For the second inequality observe that $\mathrm{span}_k^f(G, \mathbf{1}) = k$ if $\chi(G) \leq k$. If $\chi(G) > k$, then we can partition G into $\chi(G) = \chi$ color classes. With the $\binom{\chi}{k}$ k-colorable subgraphs each consisting of a different set of k color classes, we can cover every node $\binom{\chi-1}{k-1}$ times. Therefore, we have $\mathrm{span}_k^f(G, \mathbf{x}) \leq k\binom{\chi}{k}/\binom{\chi-1}{k-1} = \chi$, and so $\mathrm{span}_k^f(G, \mathbf{1})/k \leq \chi(G)/k$. □

Observe that part (b) extends the result noted earlier that G is k-perfect if $\chi(G) \leq k$. It follows directly from the definition of the fractional k-span that

$$\frac{1}{k}\ \mathrm{span}_k^f(G, \mathbf{x}) \geq \frac{1}{k+1}\ \mathrm{span}_{k+1}^f(G, \mathbf{x}).$$

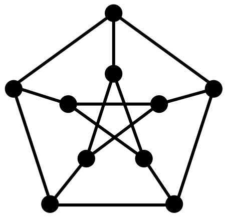


FIG. 3.1. *The Petersen graph.*

Also, if $k \geq \omega(G)$, then $\mathrm{imp}_k(G) \geq \mathrm{span}_k^f(G, \mathbf{1})/\omega_k^f(G, \mathbf{1}) = \mathrm{span}_k^f(G, \mathbf{1})/k$. Hence from part (b) above we obtain the following result.

LEMMA 3.2. *If $k \geq \omega(G)$, then $\mathrm{imp}_k(G) = \mathrm{span}_k^f(G, \mathbf{1})/k$, and $\mathrm{imp}_k(G) \geq \mathrm{imp}_{k+1}(G)$.*

In particular, the last result implies that for any $k \geq 2$ and for any triangle-free graph G, we have $\mathrm{imp}_k(G) = \mathrm{span}_k^f(G, \mathbf{1})/k$. The case $k = 1$ is different [7]: if G is a triangle-free graph which contains at least one edge, then $\mathrm{imp}(G) = \chi_f(G)/2 = \mathrm{span}_1^f(G, \mathbf{1})/2$.

If $\omega(G) \leq k \leq \chi(G)$, then by Lemma 3.1 and [15]

$$\chi_f(G) \leq k\ \mathrm{imp}_k(G) \leq \chi(G) \leq (1 + \log_2 n)\chi_f(G)$$

if G has n nodes. Hence if we could prove that it is hard to approximate the chromatic number of a triangle-free graph G up to some factor $f(n) \geq (1 + \log_2 n)$, then this would show that it is hard to approximate $\mathrm{imp}_k(G)$ up to the factor $f(n)/(1+\log_2 n)$. It is NP-hard to approximate the chromatic number up to a factor of $n^{\frac{1}{7}-\varepsilon}$ for general graphs [3]. Also, it is NP-hard to determine $\chi_f(G)$ exactly for triangle-free graphs, and hence it is NP-hard to determine $\mathrm{imp}(G)$ exactly [7].

We cannot replace $\chi(G)$ by $\chi_f(G)$ in part (b) of Lemma 3.1 as we might hope, by analogy with the case $k = 1$ (recall from [7] that $\mathrm{imp}(G) \leq \chi_f(G)/2$ if G has at least one edge), as the following example shows.

Example 3.1. The Petersen graph P, shown in Figure 3.1, satisfies

$$\mathrm{imp}_2(P) = 10/7 > 5/4 = \chi_f(P)/2.$$

For, observe that P is node-transitive, and the maximal number of nodes in a bipartite induced subgraph is 7. Thus we obtain $\mathrm{span}_2^f(P, \mathbf{1}) = 20/7$ by considering the hypergraph which has a hyperedge for each 2-colorable graph and applying, for example, Proposition 1.3.4 of [21, p. 7]. But Lemma 3.2 shows that $\mathrm{imp}_2(P) = \mathrm{span}_2^f(P, \mathbf{1})/2 = 10/7$.

Lemma 3.1 also implies that if $\omega(G) \leq k < \chi_f(G)$, then $\mathrm{imp}_k(G) > 1$: the next lemma extends this result, and will be useful in the next section.

LEMMA 3.3. *For any graph G, if $\omega(G) \leq k < \chi(G)$, then G is not k-perfect.*

Proof. Since $\omega(G) \leq k$, $(1/k)\mathbf{1} \in QSTAB_k(G)$, but since $k < \chi(G)$, $(1/k)\mathbf{1} \notin STAB_k(G)$. The result now follows from (2.4). □

The next result gives a bound on the k-imperfection ratio in terms of the imperfection ratio. It will allow us to extend the known extremal results for the case $k = 1$ [8] to cover each $k \geq 1$.

THEOREM 3.4. *For any graph G,*

$$\operatorname{imp}_k(G) \leq \frac{1}{1-(1-1/k)^k}\operatorname{imp}(G) \leq \frac{1}{1-1/e}\operatorname{imp}(G).$$

Note that $(1-1/e)^{-1} < 1.582$. To prove the theorem we need one auxiliary lemma.

LEMMA 3.5. *Let G be a graph with weight vector $\mathbf{x}$. For any $\rho \geq x_{\max}/\chi_f(G,\mathbf{x})$, we have*

$$\operatorname{span}_k^f(G,\mathbf{x}) \leq \frac{k\rho}{1-(1-\rho)^k}\chi_f(G,\mathbf{x}).$$

Proof. Let $\mathbf{y}$ be an optimal feasible solution for the LP defining $\chi_f(G,\mathbf{x})$, and let $\chi_f(G,\mathbf{x}) = \gamma$. Set $y'_S = y_S/\gamma$ for each stable set S, so $\sum_S y'_S = 1$. For each k-colorable set T of nodes in G, let

$$z_T = \sum_{\substack{S_1,S_2,\dots,S_k\\ S_1\cup S_2\cup\dots\cup S_k=T}} y'_{S_1}y'_{S_2}\dots y'_{S_k}.$$

Then

$$\sum_T z_T = \sum_{S_1}\sum_{S_2}\dots\sum_{S_k} y'_{S_1}y'_{S_2}\dots y'_{S_k} = 1.$$

(Indeed, z_T is the probability that we obtain T if we form the union of k (not necessarily distinct) stable sets picked independently at random where the stable set S has probability y'_S.) For a node $v \in V(G)$, we have

$$\begin{aligned}\sum_{T\ni v} z_T &= 1 - \sum_{S_1\not\ni v}\sum_{S_2\not\ni v}\dots\sum_{S_k\not\ni v} y'_{S_1}y'_{S_2}\dots y'_{S_k}\\ &= 1 - \left(\sum_{S_1\not\ni v} y'_{S_1}\right)\left(\sum_{S_2\not\ni v} y'_{S_2}\right)\dots\left(\sum_{S_k\not\ni v} y'_{S_k}\right)\\ &= 1 - \left(\sum_{S\not\ni v} y'_S\right)^k \geq 1-(1-x_v/\gamma)^k.\end{aligned}$$

It is easily verified that the function $f(x) = (1-(1-x)^k)/x$ is decreasing for $0 \leq x \leq 1$. Hence

$$\sum_{T\ni v} z_T \geq 1-(1-x_v/\gamma)^k = \frac{x_v}{\gamma}f(x_v/\gamma) \geq \frac{x_v}{\gamma}f(\rho).$$

Therefore $\operatorname{span}_k^f(G,\mathbf{x}) \leq k\gamma/f(p) = (k\rho/1-(1-\rho)^k)\chi_f(G,\mathbf{x})$. □

Proof of Theorem 3.4. For any $k \geq \omega(G)$ we have by Lemma 3.2 that $\operatorname{imp}_k(G) \geq \operatorname{imp}_{k+1}(G)$. Therefore it suffices to consider the case $k \leq \omega(G)$. Let $\mathbf{x}$ be a weight vector such that

$$\operatorname{span}_k^f(G,\mathbf{x}) = \operatorname{imp}_k(G) \quad \text{and} \quad \omega(G,\mathbf{x}) = 1 \geq kx_{\max}.$$

Since $\chi_f(G,\mathbf{x}) \geq \omega(G,\mathbf{x}) = 1 \geq kx_{\max}$, we have $x_{\max}/\chi_f(G,\mathbf{x}) \leq 1/k$ and $\operatorname{imp}(G) \geq \chi_f(G,\mathbf{x})$. Hence by Lemma 3.5

$$\operatorname{imp}_k(G) = \operatorname{span}_k^f(G,\mathbf{x}) \leq \frac{\chi_f(G,\mathbf{x})}{1-(1-1/k)^k} \leq \frac{\operatorname{imp}(G)}{1-(1-1/k)^k}.$$

Finally, note that $(1-1/k)^k \leq e^{-1}$. □

Theorem 3.4 together with Theorem 3.1 of [8] (which says that there exists a constant c' such that for all graphs G with $n \geq 3$ nodes $\operatorname{imp}(G) \leq c' n \log\log n / \log^2 n$) implies the following extension of the latter result.

PROPOSITION 3.6. *There exists a constant c such that for each graph G with $n \geq 3$ nodes, and each positive integer k,*

$$\operatorname{imp}_k(G) \leq c\frac{n(\log\log n)}{\log^2 n}.$$

The upper bound here is at most a factor $\log\log n$ too generous; see Theorem 7.2 below. We can also extend a result from [8] concerning graphs G with bounded maximum degree.

PROPOSITION 3.7. *For each $\varepsilon > 0$, there exists a constant d_0 such that, for each positive integer k, for each $d \geq d_0$, and for each graph G with maximum degree $\Delta(G) \leq d$,*

$$\operatorname{imp}_k(G) \leq \varepsilon d.$$

This result shows that the k-imperfection ratio grows more slowly than the maximum degree. It may be proved along similar lines to the proof of Theorem 3.2 in [8].

4. Some classes of perfect graphs. Which perfect graphs are k-perfect? In this section, we first give a polyhedral characterization. We then investigate whether a nonperfect graph can be k-perfect. This question is easily answered for $k \geq 3$, since in this case there are indeed graphs which are nonperfect but are k-perfect—just take any nonperfect graph G with $\chi(G) \leq k$. We will see that this is not true for $k = 2$: the 2-perfect graphs form a subclass of the perfect graphs, and we shall see that it is in fact a proper subclass, by considering a class of perfect graphs G (the split graphs) such that G need not be k-perfect when $k \geq 2$. Finally, we consider three standard classes of perfect graphs, namely comparability graphs, line graphs of bipartite graphs, and co-comparability graphs, and show that each graph in these classes is all-perfect (that is, k-perfect for each k).

PROPOSITION 4.1. *Let G be a perfect graph with n nodes, and let k be a positive integer. Then G is k-perfect if and only if the polytope*

$$\{\mathbf{x} \geq \mathbf{0} : \omega(G, \mathbf{x}) \leq k\} \cap [0,1]^n$$

has only integral extreme points. If the polytope has a unique nonintegral extreme point $\mathbf{z}$, then $\operatorname{imp}_k(G) = \frac{1}{k}\operatorname{span}_k^f(G, \mathbf{z})$.

Proof. Let A denote the polytope in the proposition, and let B denote the convex hull of the incidence vectors of the k-colorable sets of nodes (so that $A = k\, QSTAB_k(G)$ and $B = k\, STAB_k(G)$). Then $A \supseteq B$, and by (2.4) in Theorem 2.3, G is k-perfect if and only if $A = B$. So, if G is k-perfect, then of course the extreme points of A are $0,1$-valued. For the converse, let $\mathbf{z}$ be any integral point in A. Then $\mathbf{z}$ is the incidence vector of the nodes in a subgraph H of G with $\omega(H) \leq k$, and so with $\chi(H) \leq k$: hence $\mathbf{z} \in B$. Hence if each extreme point of A is integral, then $A \subseteq B$, and it follows that G is k-perfect. This completes the proof of the first part of the proposition.

Further, it now follows using (2.5) in Theorem 2.3 that if A has a nonintegral extreme point $\mathbf{z}$, then $\operatorname{imp}_k(G)$ is the maximum value of $\frac{1}{k}\operatorname{span}_k^f(G, \mathbf{z})$ over such points $\mathbf{z}$. ∎

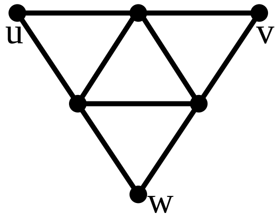


FIG. 4.1. *The "Hajos" graph G_2 as in Proposition* 4.3.

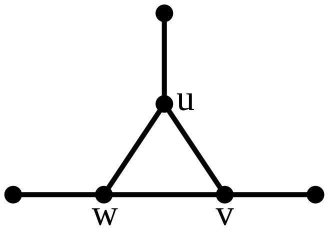


FIG. 4.2. *The complement of the "Hajos" graph.*

The next two propositions show that the 2-perfect graphs form a proper subclass of the perfect graphs.

PROPOSITION 4.2. *Each* 2*-perfect graph is perfect.*

Proof. Recall from [7] that the *binary imperfection ratio* $\text{imp}^{\text{b}}(G)$ is the maximum value of $\text{span}_1(G, \mathbf{x})/\omega(G, \mathbf{x})$ over all nonzero $0, 1$ weight vectors $\mathbf{x}$: also $\text{imp}^{\text{b}}(G) \leq \text{imp}(G)$, and $\text{imp}^{\text{b}}(G) = 1$ if and only if G is perfect. We claim that $\text{imp}_2(G) \geq \text{imp}^{\text{b}}(G)$ for each graph G. But then, if G is not perfect, we have $\text{imp}_2(G) \geq \text{imp}^{\text{b}}(G) > 1$ and so G is not 2-perfect, and the proposition follows.

To prove the claim, note first that if G consists of isolated nodes, then $\text{imp}_2(G) = 1 = \text{imp}^{\text{b}}(G)$, so we may assume that G has at least one edge. Let $\mathbf{x}$ be a 0,1 weight vector of G with $\text{imp}^{\text{b}}(G) = \chi_f(G, \mathbf{x})/\omega(G, \mathbf{x})$ and $\omega(G, \mathbf{x}) \geq 2$. Then $\omega_2^f(G, \mathbf{x}) = \omega(G, \mathbf{x})$, and hence

$$\text{imp}_2(G) \geq \frac{\text{span}_2^f(G, \mathbf{x})}{\omega_2^f(G, \mathbf{x})} \geq \frac{\chi_f(G, \mathbf{x})}{\omega(G, \mathbf{x})} = \text{imp}^{\text{b}}(G)$$

as claimed. □

The next proposition shows that for each $k \geq 2$ there are perfect graphs which are not k-perfect. Recall that a *split graph* is a graph the nodes of which can be covered by a clique and a stable set. It is well known and easy to see that such graphs are perfect; see, for example, [12].

PROPOSITION 4.3. *For each $k \geq 2$, there exists a split graph G_k which is not k-perfect.*

Proof. Consider the graph G_k which consists of a clique of size $2k - 1$ and a stable set of size $\binom{2k-1}{k}$ such that every k-subset of nodes of the clique is adjacent to exactly one node of the stable set. For the graph G_2, the "Hajos graph," see Figure 4.1. Let $\mathbf{x}^k$ be the weighting of G_k with $x_u^k = 1$ for each node u of the clique, and $x_v^k = 2$ for each node v of the stable set. Since $k \geq 2$, we have $\omega_k^f(G_k, \mathbf{x}^k) = 2k$ but $\text{span}_k^f(G_k, \mathbf{x}^k) > 2k$. For suppose that there is a solution $\mathbf{y}$ to the LP for $\text{span}_k^f(G_k, \mathbf{x}^k)$ with value $2k$. Then any k-colorable graph S actually used (that is, with $y_S > 0$) must contain all the nodes of the stable set, and so can contain at most $k - 1$ nodes of the clique. Hence the total weight covered on the nodes of the clique is at most $2k - 2 < 2k - 1$, and so the covering is not feasible for G_k and $\mathbf{x}^k$. In summary, $\text{imp}_k(G_k) \geq \text{span}_k^f(G_k, \mathbf{x}^k)/\omega_k^f(G_k, \mathbf{x}^k) > 1$. □

Example 4.1. Let us consider more carefully the Hajos graph H shown in Figure 4.1, which we have already noted is perfect. We shall see that H is a minimal non-2-perfect graph and that $\text{imp}_2(H) = \frac{9}{8}$.

It is easy to check that any proper induced subgraph of H is an interval graph, and hence it is a co-comparability graph and thus is all-perfect by Proposition 4.8

below. Let $\mathbf{z}$ be the demand vector with $z_v = 1$ on the three degree 2 nodes and $z_v = \frac{1}{2}$ on the other three nodes. We claim that $\mathbf{z}$ is the unique nonintegral vertex of $2\,QSTAB_2(H)$. Then by Proposition 4.1, $\mathrm{imp}_2(H) = \frac{1}{2}\mathrm{span}_2(H, \mathbf{z})$. Since the maximum number of nodes in a bipartite subgraph of H is 4, and $\sum_v z_v = \frac{9}{2}$, it follows that $\mathrm{span}_2(H, \mathbf{z}) \geq \frac{9}{4}$. But it is straightforward to find an appropriate covering which shows that equality holds.

It remains to establish the claim. Let $\mathbf{x}$ be a nonintegral vertex of $2\,QTAB_2(H) = 2\,QSTAB(H) \cap [0,1]^V$. Since each proper subgraph of H is 2-perfect, we must have that each $x_v > 0$. Also, since each vertex corresponds to a basic feasible solution and H has 6 nodes, there must be at least 6 tight constraints.

Suppose that all 4 triangles yield a tight constraint. Then opposite pairs of nodes must have the same value x_v. (An opposite pair consists of a degree-2 node and the nonadjacent degree-3 node.) Also, at least 6–4=2 coordinates x_v equal 1. Let the values on the opposite pairs be 1, x, and y, where $0 < x \leq y < 1$. (Note that $y < 1$ since $x + y \leq 1$.) But then $\mathbf{x}$ is not a vertex, since we could replace x, y by $x \pm \delta$ and $y \mp \delta$, where $\delta = \min\{x, 1-y\} > 0$.

Hence at most 3 triangles yield a tight constraint, and so $x_v = 1$ for at least 3 nodes v. But no two of these nodes can lie on a triangle, so they must be the three degree-2 nodes. Now we are forced to put value $\frac{1}{2}$ on the other nodes. Thus indeed $\mathbf{z}$ is the unique nonintegral vertex of $2\,QSTAB(H) \cap [0,1]^V$, as claimed.

Now we consider three classes of perfect graphs G such that each G is all-perfect, namely comparability graphs, line graphs of bipartite graphs, and co-comparability graphs. Before we proceed further, let us remark that in contrast to the case $k = 1$, when $k \geq 2$ the complement $\overline{G}$ of a k-perfect graph G need not be k-perfect. Consider, for example, the odd holes and antiholes; see Proposition 5.3 below. Also, the Hajos graph G_2 shown in Figure 4.1 is not 2-perfect, but its complement shown in Figure 4.2 is the line-graph of a bipartite graph and so is 2-perfect, indeed all-perfect; see Proposition 4.7 below.

A graph is a *comparability graph* if there exists a partial order of the nodes such that distinct nodes u and v are adjacent exactly when they are comparable in the partial order. We use one lemma to prove that comparability graphs are all-perfect (and indeed we use this lemma again in section 6). The lemma involves "circular" (or "cyclic") interval colorings. An *m-circular interval coloring* of G with integral weight vector $\mathbf{x}$ is a multicoloring of the nodes of G using the colors $0, 1, \ldots, m-1$ such that for each node v of G, the set $\{i : v \text{ has color i}\}$ has cardinality x_v and forms an interval in the cyclic order $(0, 1, \ldots, m-1)$.

LEMMA 4.4. *Let G be a graph with integral weight vector $\mathbf{x}$. Suppose that there is an m-circular interval coloring of the graph $G, \mathbf{x}$, where m satisfies $m \geq kx_{\max}$ and $m \equiv 1\ (mod k)$. Then* $\mathrm{span}_k(G, \mathbf{x}) \leq m$.

Proof. Consider the assignment $\phi(v) = \{ki (\mathrm{mod}\, m) : v \text{ has color } i\}$. Then for $i, j \in \{0, 1, \ldots, m-1\}$, $ki = kj$ is equivalent to $i = j$ since $m \equiv 1 (\mathrm{mod}\, k)$. It follows that $|\phi(v)| = x_v$ for each node v and $\phi(u) \cap \phi(v) = \emptyset$ for adjacent nodes u and v.

It remains to show that for each node $v \in V(G)$ and any two distinct elements $c_1, c_2 \in \phi(v)$, we have $|c_1 - c_2| \geq k$. For each node $v \in V(G)$, two distinct elements of $\phi(v)$ are of the form $kc + ik (\mathrm{mod}\, m)$ and $kc + jk (\mathrm{mod}\, m)$ with $0 \leq i < j \leq x_v - 1$ by the definition of ϕ and the fact that the colors of v form an interval in the cyclic order $0, 1, \ldots, m-1$. But $|k(j-i) - 0| = k(j-i) \geq k$ and $|m - k(j-i)| \geq m - (x_{\max} - 1)k \geq k$, and the assignment is therefore k-feasible and uses only the colors $0, \ldots, m-1$. □

PROPOSITION 4.5. *Each comparability graph is all-perfect.*

Proof. Let k be a positive integer, and let $G = (V, E)$ be a comparability graph. Let $\prec$ be a partial order on V such that distinct u and v are comparable if and only if $\{u, v\}$ is an edge of G. Let D be the corresponding acyclic (transitive) orientation of G, where we orient the edge $\{u, v\}$ from u to v if $u \prec v$. Let $\mathbf{x}$ be a integral weight vector. Form an acyclic directed graph D' from D by replacing each node v by a directed path of x_v nodes. Thus D' has nodes $v_1, \ldots, v_{x_v}$ for each $v \in V$, and there is an arc u_iv_j in D' if and only if either $u = v$ and $i < j$, or $u \neq v$ and uv is an arc in D.

For each node v_i in D', let $\phi(v_i)$ be the maximum length (i.e., number of arcs) in a path in D' ending at v_i. Then ϕ takes values in $\{0, 1, \ldots, \omega(G, \mathbf{x}) - 1\}$; for each node $v \in V$, ϕ takes distinct consecutive values on the x_v nodes $v_1, \ldots, v_{x_v}$ of D'; and if uv is an arc of D, then $\phi(u_i) < \phi(v_j)$ for each i, j. Thus ϕ gives a proper coloring of $(G, \mathbf{x})$, using colors $\{0, 1, \ldots, \omega(G, \mathbf{x}) - 1\}$, such that for each node $v \in V$ the colors on v are consecutive. Hence there is an m-cyclic interval coloring of $G, \mathbf{x}$ with $m \leq \omega_k(G, \mathbf{x}) + k - 1$. So by Lemma 4.4, $\text{span}_k(G, \mathbf{x}) \leq \omega_k(G, \mathbf{x}) + k - 1$, and the result now follows by (2.6). □

The first part of the above proof is not new. We defined circular interval colorings above, and in a similar way, when we use ordinary linear channels, we may define an *interval coloring* of $G, \mathbf{x}$. The *interval span* $\text{ispan}(G, \mathbf{x})$ is the smallest number for which an interval coloring exists. A graph G is called *superperfect* in [10] if $\text{ispan}(G, \mathbf{x}) = \omega(G, \mathbf{x})$ for each integral weight vector $\mathbf{x}$. Observe that by Lemma 4.4, any superperfect graph is all-perfect. Hoffman showed that any comparability graph is superperfect; see [10]. It is also shown there that any graph in a certain class is superperfect, where this class contains the complements of the even cycles.

We shall consider two more classes of perfect graphs and show that each graph in these classes is all-perfect. These classes are the line-graphs of bipartite graphs and the co-comparability graphs (that is, complements of comparability graphs). We give a unified proof treatment, based on the polyhedral characterization in Proposition 4.1.

Recall that a matrix is *totally unimodular* if each square submatrix has determinant 0 or ±1. Let us call a polyhedron *totally unimodular* if it may be expressed as $\{\mathbf{x} : A\mathbf{x} \leq \mathbf{b}\}$ for some totally unimodular matrix A and integral vector $\mathbf{b}$. It is well known that such a polyhedron is *integral*; that is, it has the property that each face contains an integral vector; see, for example, [20]. If $S \subseteq R^n$ and I is a nonempty subset of the indices $\{1, \ldots, n\}$, we call $\{\mathbf{x} \in R^I : (\mathbf{x}, \mathbf{y}) \in S \text{ for some } \mathbf{y}\}$ the *projection* of S onto the coordinates I. If we start with an integral polyhedron and project onto some set of coordinates, then the resulting polyhedron is again integral.

LEMMA 4.6. *Let G be a perfect graph, and let $P = STAB(G)$ $(= QSTAB(G))$. If P is a totally unimodular polyhedron, or more generally the projection of such a polyhedron onto some set of coordinates, then G is all-perfect.*

Proof. Note first that if the $(m \times n)$ matrix A is totally unimodular, then so is any submatrix of the $((m + 2n) \times n)$ matrix obtained by stacking the matrices $A, I_n, -I_n$ above one another (where I_n denotes the $(n \times n)$ identity matrix). It follows that if P satisfies the condition in the lemma, then so does $Q = kP \cap [0, 1]^V$ for any positive integer k. Then Q is the projection of an integral polyhedron onto some set of co-ordinates, and so Q is integral. Hence the result follows from Proposition 4.1. □

Now let us consider the line-graphs of bipartite graphs and use Lemma 4.6 to show that such graphs are all-perfect. The complements of these graphs need not be all-perfect: we have already seen that the Hajos graph is not 2-perfect, and it is the complement of the line-graph of a bipartite graph.

PROPOSITION 4.7. *Let $G = (V, E)$ be a bipartite graph. Then the line-graph $L(G)$ is all-perfect.*

Proof. Observe that

$$QSTAB(L(G)) = \left\{ \mathbf{x} \in R_+^E : \sum_{e: v \in e} x_e \leq 1 \ (\forall v \in V) \right\},$$

which is a totally unimodular system (see, for example, [20]); so the result follows from Lemma 4.6. □

Finally in this section, we use Lemma 4.6 to show that co-comparability graphs are all-perfect.

PROPOSITION 4.8. *Each co-comparability graph is all-perfect.*

Proof. Let $\prec$ be a partial order on V such that distinct nodes u and v are adjacent in G if and only if they are incomparable under $\prec$. We construct a directed graph D as follows. There are nodes v^- and v^+ for each node v in V, together with a new source node s and sink node t. There is an arc st, there are arcs sv^- and v^+t for each $v \in V$, and there are arcs u^+v^- for each pair of nodes $u, v \in V$ with $u \prec v$. Also, there is an arc v^-v^+ for each $v \in V$. We shall identify in the obvious way a vector indexed by the arcs v^-v^+ with a vector indexed by V.

Note that a stable set in G corresponds to an $s - t$ path in D, and a convex combination of incidence vectors of stable sets of G corresponds to a unit volume $s-t$ flow in D. Now $\mathbf{x} \in STAB(G)$ if and only if $\mathbf{x}$ is a convex combination of incidence vectors of stable sets of G. Thus $\mathbf{x} \in STAB(G)$ if and only if $\mathbf{x}$ "is" the projection onto the arcs v^-v^+ of a unit volume $s - t$ flow in D. But such flows in D form a totally unimodular polyhedron (since the node-arc incidence matrix of D is totally unimodular), so the result follows from Lemma 4.6. □

5. Minimal non-k-perfect graphs. In this section we consider minimal non-k-perfect graphs, in other words, graphs G which are not k-perfect, but deleting any node yields a k-perfect graph. The strong perfect graph theorem [2] asserts that the only minimal non-1-perfect graphs are the odd holes and antiholes, that is, the odd cycles C_n for $n \geq 5$ and their complements; so there would be a "small" list of excluded induced subgraphs for perfection. Can we hope for such a concise result for k-perfect graphs? Regrettably, the answer is no, at least not in this form; see Theorem 5.4.

Before we prove this theorem, we consider the k-imperfection ratio of minimal non-k-perfect graphs and of the odd holes and antiholes. We show first that for any minimal non-k-perfect graph G on n nodes, $\operatorname{imp}_k(G) \leq n/(n-1)$. To do this we need the following lemma.

LEMMA 5.1. *If the nodes of a graph G can be covered q times by p induced subgraphs $H_1, \ldots, H_p$, then*

$$\operatorname{imp}_k(G) \leq \frac{1}{q} \sum_{i=1}^{p} \operatorname{imp}_k(H_i).$$

Proof. For every weight vector $\mathbf{x}$ of G, we have

$$q \operatorname{span}_k^f(G, \mathbf{x}) = \operatorname{span}_k^f(G, q\mathbf{x}) \leq \sum_{i=1}^{p} \operatorname{span}_k^f(H_i, \mathbf{x})$$

$$\leq \sum_{i=1}^{p} \operatorname{imp}_k(H_i)\, \omega_k^f(H_i, \mathbf{x}) \leq \omega_k^f(G, \mathbf{x}) \sum_{i=1}^{p} \operatorname{imp}_k(H_i),$$

and the result now follows by the definition of $\operatorname{imp}_k(G)$. □

PROPOSITION 5.2. *For each $k \geq 1$, if G is a minimal non-k-perfect graph on n nodes, then*

$$\operatorname{imp}_k(G) \leq \frac{n}{n-1}.$$

Proof. The removal of any node v yields a k-perfect graph, and hence G can be covered $n-1$ times by n k-perfect graphs. Lemma 5.1 now yields the result. □

We can now determine the k-imperfection ratios of the odd holes and antiholes. Recall that we already know that even cycles and their complements are k-perfect for all k, since even cycles are bipartite, and thus are comparability graphs. Let n be odd and at least 5. Then C_n is k-perfect for $k \geq 3$, since $\chi(C_n) \leq 3$, and $\overline{C}_n$ is k-perfect for all $k \geq (n+1)/2$ since $\chi(\overline{C}_n) \leq (n+1)/2$. The following proposition completes the picture.

PROPOSITION 5.3. *Let $n \geq 5$ be an odd integer. Then the odd hole C_n is minimal non-2-perfect, and* $\operatorname{imp}_2(C_n) = \frac{n}{n-1}$. *Also, for each $k = 1, \ldots, \frac{n-1}{2}$, the odd antihole $\overline{C}_n$ is minimal non-k-perfect, and* $\operatorname{imp}_k(\overline{C}_n) = \frac{n}{n-1}$.

Proof. Since $\omega(C_n) = 2$ and $\chi_f(C_n) = \frac{2n}{n-1}$, Lemma 3.1 shows that $\operatorname{imp}_k(C_n) \geq \frac{n}{n-1}$. Since bipartite graphs are 2-perfect, it follows that C_n is minimal non-2-perfect, and so $\operatorname{imp}_k(C_n) \leq \frac{n}{n-1}$ by the last proposition.

Since $\omega(\overline{C_n}) = \frac{n-1}{2}$ and $\chi_f(\overline{C_n}) = \frac{n}{2}$, Lemma 3.1 shows that $\operatorname{imp}_k(\overline{C_n}) \geq \frac{n}{n-1}$. Since co-bipartite graphs are 2-perfect, it follows that $\overline{C_n}$ is minimal non-k-perfect, and so $\operatorname{imp}_k(\overline{C_n}) \leq \frac{n}{n-1}$ by the last proposition. □

We saw at the end of the introduction that cliques and stable sets always form k-perfect graphs. Hence by Lemma 5.1, the cochromatic number $z(G)$ of G is an upper bound on $\operatorname{imp}_k(G)$. (Recall that the cochromatic number $z(G)$ is the least number of stable sets and cliques needed to cover the graph G.) For the case $k = 1$, one can strengthen this bound and obtain $\operatorname{imp}_k(G) \leq z(G)/2$ for any nontrivial graph G [7]. Proposition 4.3 shows that this is not true when $k \geq 2$, but it is easy to see that $\operatorname{imp}_k(G) < z(G)/2 \ + 1$ by noting that any two stable sets and any two cliques induce an all-perfect graph.

We now consider the number of (node-)minimal non-k-perfect graphs on n nodes and show that when $k \geq 2$ there are many such graphs.

THEOREM 5.4. *For each integer $k \geq 2$, let $f_k(n)$ be the number of nonisomorphic minimal non-k-perfect graphs on at most n nodes. Then $f_k(n)$ grows at least exponentially with n.*

The rest of this section is devoted to proving this result. We first consider the easier case when $k \geq 3$. After that, we start the proof for the case $k = 2$, then break to state and prove four lemmas, and then complete the proof.

Proof. Consider first the case $k \geq 3$. We call a graph G k-critical if $\chi(G) = k$ and deleting any node yields a graph with chromatic number $k - 1$. Note that each $(k+1)$-critical graph G other than K_{k+1} has $\omega(G) \leq k$, and so G is node-minimal non-k-perfect by Proposition 3.3. It is known [23] that the number of nonisomorphic 4-critical graphs on at most n nodes is at least $c^{(n^2)}$ for some $c > 1$. By adding to a 4-critical graph G a clique with l nodes, each of which is adjacent to each of the nodes of G, one obtains a $(4+l)$-critical graph. This completes the proof for the case

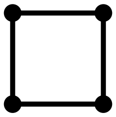


FIG. 5.1. *The seed graph H.*

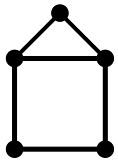


FIG. 5.2. *The graph H'.*

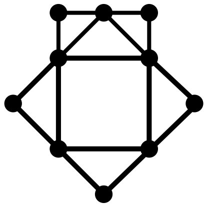


FIG. 5.3. *The graph $G = G(H, e)$.*

$k \geq 3$, but this approach will not work for $k = 2$, as the only 3-critical graphs are the odd cycles.

We now consider the case $k = 2$. We shall show that each graph G formed as below is node-minimal non-2-perfect (and outerplanar and perfect). The smallest graph G we shall construct is the Hajos graph, as shown in Figure 4.1.

Construct the graph G as follows. Start with any 2-node-connected outerplanar bipartite graph H with at least one edge, the "seed" graph. It has a unique outerplanar embedding with each node on the infinite face. Pick an edge $e = uv$ on the infinite face and add a new degree-2 node v^* adjacent to u and v. The new graph H' is 2-node-connected and outerplanar, with a unique outerplanar embedding such that each node is on the infinite face. Let C be the Hamilton cycle bounding the infinite face, which is in fact the unique Hamilton cycle in H. Note that C has an odd number of nodes. For each edge $f = ab$ on C, add a new degree-2 node v_f adjacent to a and b. This gives the desired graph $G = G(H, e)$; see Figures 5.1–5.3. The graph G is outerplanar, and as it has no odd holes, it is perfect [24]. It remains to show three things.

1. It is easy to see that the number of nonisomorphic graphs G as above on at most n nodes grows at least exponentially with n, since this holds for the seed graphs H.
2. The graph G is not 2-perfect. For we may give weight 2 to each of the degree-2 nodes added at the last step when we formed G from H', and weight 1 to each of the other nodes (which came from H'). It is easy to see that $\omega_2(G, \mathbf{x}) = 4$. But $\mathrm{span}_2^f(G, \mathbf{x}) > 4$. For suppose that there is a solution $\mathbf{y}$ to the LP for $\mathrm{span}_2^f(G, \mathbf{x})$ with value 4. Then any 2-colorable graph S actually used (that is, with $y_S > 0$) must contain all the nodes v with $x_v = 2$ and so can contain at most $\frac{|C|-1}{2}$ nodes of the circuit C. Hence the total weight covered on the nodes of C is at most $|C| - 1$, and so the covering is not feasible for $G, \mathbf{x}$.
3. Finally, we must check that any proper induced subgraph of G is 2-perfect. We shall complete this last requirement, and thus complete the proof of the theorem, after stating and proving four lemmas.

The next lemma (together with Proposition 1.1) shows that each node-minimal non-k-perfect graph is 2-node-connected.

LEMMA 5.5. *Let the graph G be connected, with a cut node v. Let $G_1, G_2, \ldots$ be the components formed when the node v is deleted, and for $i = 1, 2, \ldots$ let H_i be the graph formed by adding back the node v to G_i (that is, H_i is the subgraph of G induced by $V(H_i) \cup \{v\}$). Then for any positive integer k,*

$$\mathrm{imp}_k(G) = \max_i \{\mathrm{imp}_k(H_i)\}.$$

Proof. It suffices to note that if $A_i \subseteq V(H_i)$ is k-colorable for each $i = 1, 2, \ldots$

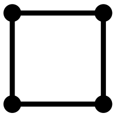


FIG. 5.4. *A bipartite graph H.*

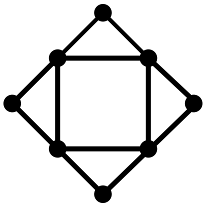


FIG. 5.5. *The graph $\hat{H}$.*

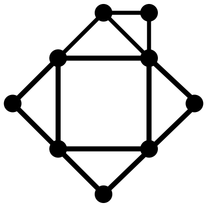


FIG. 5.6. *The all-perfect graph G.*

(that is, each induced subgraph $G[A_i]$ is k-colorable) and either $v \in \cap_i A_i$ or $v \notin \cup_i A_i$, then $\cup_i A_i$ is k-colorable. □

LEMMA 5.6. *Let the* 2*-node-connected graph G have no odd holes. Suppose that there is a separating set consisting of a node u and an edge vw, where it is not the case that u is adjacent to both v and w. Let G_1 and G_2 be the components formed when the separating set is deleted, and let H_1 and H_2 be the graphs formed by adding back the node u to G_1 and G_2, respectively. Then*

$$\text{imp}_2(G) = \max\{\text{imp}_2(H_1), \text{imp}_2(H_2)\}.$$

Proof. It suffices to show that if $A_1 \subseteq V(H_1)$ and $A_2 \subseteq V(H_2)$ are 2-colorable and either $u \in A_1 \cap A_2$ or $u \notin A_1 \cup A_2$, then $A_1 \cup A_2$ is 2-colorable. This is obvious if $u \notin A_1 \cup A_2$, so assume that $u \in A_1 \cap A_2$. Suppose that there is an odd cycle contained in $A_1 \cup A_2$. This cycle must go through both the node u and the edge vw. Without loss of generality, we may assume that v is in H_1 and w is in H_2. There is a u-v path in H_1: consider a shortest such path Q_1. Similarly, there is a u-w path in H_2: consider a shortest such path Q_2. Then the cycle formed from Q_1, Q_2, and the edge vw is an odd hole in G, a contradiction. □

LEMMA 5.7. (a) *Start with a bipartite graph H. For each edge $e = ab$, add a new degree-*2 *node v_e adjacent to a and b. Then the graph $\hat{H}$ formed is all-perfect.*

(b) *Now take an edge a_0b_0 in $\hat{H}$, where the node a_0 is in H and the node b_0 is in $\hat{H}$ but not in H, and add a new degree-*2 *node v^* adjacent to a_0 and b_0. Then the graph G formed is all-perfect.*

See Figures 5.4–5.6 for an illustration of the construction.

Proof. It suffices to prove (b). Let $\mathbf{x}$ be a weight vector for G. Denote $\omega(G, \mathbf{x})$ by ω. We shall show that there is an interval coloring of $G, \mathbf{x}$ using colors $1, \ldots, \omega$. The result will then follow by Lemma 4.4.

Properly color the nodes of H with the two labels "low" and "high," where a_0 is "low." Give each "low" node v the "low" colors $1, \ldots, x_v$; give each "high" node v the "high" colors $\omega - x_v + 1, \ldots, \omega$. For each edge e of H, assign node v_e in $\hat{H}$ the interval $x_a + 1, \ldots, x_a + x_{v_e}$, where a is the "low" node incident with e. This gives an interval coloring of $\hat{H}, \mathbf{x}$ with colors $1, \ldots, \omega$.

Finally we handle the node v^* formed at the last stage. Note that node a_0 has been assigned the "low" interval of colors $1, \ldots, x_{a_0}$, and node b_0 has been assigned the "next" interval of colors $x_{a_0} + 1, \ldots, x_{a_0} + x_{b_0}$. Thus we may assign to v^* the "high" interval of colors $\omega - x_{v^*} + 1, \ldots, \omega$. □

Completion of the proof of Theorem 5.4. Recall that we must check that any proper induced subgraph of the graph $G = G(H, e)$ is 2-perfect. We use induction on the number of nodes of the seed graph H. The base case is the Hajos graph, which we have already handled. Now let $G = G(H, e)$, where H has Hamilton circuit C, and suppose that we know the result for any smaller seed graph. Let v be a node in G, and let G' be the graph $G-v$ obtained from G by deleting v. We must show that G' is 2-perfect. Let T denote the unique triangle in H'. We consider four cases. The first two cover the possibilities when v is in H', and the second two when v is not in H' (and so v has degree 2).

1. Suppose that v is in T. Then $H'-v$ is bipartite, so G' is 2-perfect by Lemma 5.7(a).
2. Suppose that v is in H' and not in T, and so v is in H. Consider the outerplanar embedding of H'. Let F be a bounded face such that its boundary cycle D contains v. Let x be a node on D not adjacent to v. Then x is a cutnode for G'. Let $\tilde{G}$ be the graph obtained by adding x back to the component of $G'-v$ which contains v^*. Then $\mathrm{imp}_2(G') = \mathrm{imp}_2(\tilde{G})$ by Lemmas 5.5 and 5.7(a). But $\tilde{G}$ is 2-perfect by Lemma 5.5 and the induction hypothesis.
3. Suppose that v is v_f for some edge $f=ab$ in C and not in T. Consider the outerplanar embedding of H'. Let F be the bounded face containing f on its boundary. There is a node x on F other than a and b such that x and the edge f form a separating set S for G'. Let $\tilde{G}$ be the graph obtained by adding x back to the component of $G'-S$ which contains v^*. Then $\mathrm{imp}_2(G') = \mathrm{imp}_2(\tilde{G})$ by Lemmas 5.6 and 5.7(a). But $\tilde{G}$ is 2-perfect by Lemma 5.5 and the induction hypothesis.
4. The remaining case is when v is v_f for one of the two edges f of C in T, and this is exactly the case covered by Lemma 5.7(b). □

6. Disk graphs. In this section we bound the k-imperfection ratio of unit disk graphs, general disk graphs, and induced subgraphs of the triangular lattice (which are a subclass of unit disk graphs).

In a *unit disk graph* the nodes can be represented by unit diameter (closed) disks in the plane such that two distinct nodes are adjacent if and only if the corresponding disks intersect. These graphs are important in radio channel assignment, since we obtain a unit disk graph as an interference graph if we assume that the service area of a transmitter corresponds to a unit size disk. It is known [21, pp. 60–63] that we can fractionally cover the nodes of a unit disk graph G d times by about $4.36d$ graphs, which are disjoint unions of cliques, and hence by Lemma 5.1 we have $\mathrm{imp}_k(G) \leq 4.36$. The next result improves this bound: it extends Proposition 3.3 of [7], which is the case $k=1$.

PROPOSITION 6.1. *For each unit disk graph G and each positive integer k,*

$$\mathrm{imp}_k(G) \leq 1 + 2/\sqrt{3} \sim 2.155.$$

Proof. If the center of each disk lies in a stripe of width $\sqrt{3}/2$, then the corresponding unit disk graph is a co-comparability graph [11], and so is all-perfect by Proposition 4.8. If t is sufficiently large, then with t such graphs we can cover a given finite unit disk graph at least $\frac{\sqrt{3}}{2+\sqrt{3}}\,t$ times; see the proof of Proposition 3.3 of [7]. The result now follows by Lemma 5.1. □

A generalization of a unit disk graph is a disk graph. A *disk graph* is a graph the nodes of which can be represented by (closed) disks in the plane such that two nodes are adjacent if and only if the corresponding disks intersect. (The nodes may correspond to transmitters with different powers.) It is easy to verify that the neighborhood of the node represented by a smallest size disk can be covered by 6 cliques. Hence the bound below follows from the lemma after it.

PROPOSITION 6.2. *For each disk graph G and each positive integer k,*

$$\operatorname{imp}_k(G) \le \begin{cases} 6 & \text{if } \ k \le 6, \\ 6 - \frac{6}{k} & \text{if } \ k \ge 6. \end{cases}$$

LEMMA 6.3. *For each graph G and $t \ge 1$, if each induced subgraph of G contains a node the neighborhood of which can be covered at least p/t times by a family of p cliques, then* $\operatorname{imp}_k(G) \le t + \max\{0, 1 - t/k\} < t + 1$.

Proof. Let G have n nodes. We can order the nodes of G in such a way that, for each $i = 2, \ldots, n$, the nodes of $\{v_1, \ldots, v_{i-1}\}$ which are adjacent to v_i can be covered q_i times by a family of p_i cliques, where $p_i/q_i \le t$. Consider any integral weight vector $\mathbf{x}$ for G. Now, we greedily color the nodes of G in the order above, i.e., when we come to color the node v we assign to it the lowest color, say c, which is not already assigned to a neighbor of v, then the lowest color available which is at least $c + k$, and so on until the node v is colored x_v times. Clearly, we obtain a k-feasible assignment. We claim that this procedure uses only the colors up to $t\omega_k(G, \mathbf{x}) + (t - k)$ if $k \le t$ and the colors up to $(t + 1 - t/k)\omega_k(G, \mathbf{x})$ if $t < k$. To see this, for each $i = 2, \ldots, n$, let $w(i)$ be the sum of the values x_{v_j} over all neighbors v_j of v_i with $j < i$. Observe that we have $p_i(\omega(G, \mathbf{x}) - x_{v_i}) \ge q_i w(i)$, and so $w(i) \le t(\omega(G, \mathbf{x}) - x_{v_i})$. When we come to color v_i, at most $w(i)$ colors are already used for the neighbors of v_i. Thus we can color v_i with x_{v_i} colors using only colors up to $t(\omega(G, \mathbf{x}) - x_{v_i}) + k(x_{v_i} - 1) = t\omega(G, \mathbf{x}) + (k - t)x_{v_i} - k$. Therefore we use only the colors up to $t\omega(G, \mathbf{x}) - k \le t\omega_k(G, \mathbf{x}) + (t - k)$ if $k - t \le 0$, and only the colors up to $t(\omega(G, \mathbf{x}) - 1) + (1 - t/k)(x_{\max} - 1)k \le (t + 1 - t/k)\omega_k(G, \mathbf{x})$ if $k - t > 0$. The result now follows by (2.6). ▯

A subclass of unit disk graphs, the class of finite induced subgraphs G of the triangular lattice, has attracted considerable attention from researchers interested in the channel assignment problem. The reason for this interest is the fact that when the potential service area for each transmitter is a unit diameter disk in the plane, arranging the transmitters on a triangular lattice is most efficient, in the sense of achieving universal coverage with as few transmitters as possible. Observe that such a graph G is k-perfect for each $k \ge 3$, since $\chi(G) \le 3$.

PROPOSITION 6.4. *Let G be a finite induced subgraph of the triangular lattice. Then* $\operatorname{imp}_k(G) \le \frac{4}{3}$ *for $k = 1$ and $k = 2$.*

Proof. For $k = 1$ we can use a result obtained in [18], which says that $\operatorname{span}_1(G, \mathbf{x}) \le (4\omega_1(G, \mathbf{x}) + 1)/3$ and thus implies that $\operatorname{imp}(G) \le 4/3$ by (2.6); see also [7]. To prove the result for $k = 2$ consider a weight vector $\mathbf{x}$ of G. We will show that

$$\operatorname{span}_2(G, \mathbf{x}) \le \frac{4\omega_2(G, \mathbf{x}) + 8}{3}. \tag{6.1}$$

The result will then follow by (2.6). This result appears in [22] in a slightly weaker form, but since the improvement is easy from [18] once you know Lemma 4.4, we spell out a proof.

We may assume without loss of generality that $\omega(G, \mathbf{x}) \ge 2x_{\max}$, since for graphs with no edges the result is trivial, and for all other graphs if $\omega(G, \mathbf{x})$ is strictly less than $2x_{\max}$, then we could increase the weights at some nodes without increasing $\omega_2(G, \mathbf{x})$. Let us denote $\omega(G, \mathbf{x})$ simply by ω. By the proof of the main result in [18] we can find weight vectors $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ such that $\mathbf{x} = \mathbf{x}^{(1)} + \mathbf{x}^{(2)}$, and the following holds:

- there is an ω-cyclic interval coloring of $G, \mathbf{x}^{(1)}$, and
- the subgraph H of G induced by the nodes v with $x_v^{(2)} > 0$ is bipartite (indeed acyclic), and $x_v^{(2)} \le (\omega + 2)/6$ for each v.

Hence by Lemma 4.4, $\text{span}_2(G, \mathbf{x}^{(1)}) \leq \omega + 1$, and since the graph H is bipartite, $\text{span}_2(G, \mathbf{x}^{(2)}) \leq 2x^{(2)}_{\max} \leq (\omega + 2)/3$. Hence

$$\text{span}_2(G, \mathbf{x}) \leq \text{span}_2(G, \mathbf{x}^{(1)}) + \text{span}_2(G, \mathbf{x}^{(2)}) + 1 \leq (4\omega + 8)/3,$$

as required. □

For $k = 1$ and $k = 2$ we can have $\text{imp}_k(G) > 1$ for a finite induced subgraph of the triangular lattice, since the cycle C_9 on 9 nodes is such a graph and we have already seen that $\text{imp}(C_9) = \text{imp}_2(C_9) = 9/8$. It would be interesting to determine whether in fact $\text{imp}_2(G) \leq 9/8$ for all induced subgraphs of the triangular lattice; see [7] for a discussion on the corresponding question for $\text{imp}(G)$.

7. Random graphs.

In this section we use results on the imperfection ratio of a random graph from [8] to prove corresponding results for the k-imperfection ratio. First we need one deterministic lemma, which gives a bound on the k-imperfection ratio of a graph G in terms of the imperfection ratios of k induced subgraphs of G.

LEMMA 7.1. *Let $V_0, \dots, V_{k-1}$ be a partition of the node set of a graph G, and let G^i be the subgraph induced by V_i for $i = 0, \dots, k-1$. Then*

$$\text{imp}_k(G) \leq k \max_i \{\text{imp}(G^i)\},$$

where the maximum is over $i = 0, \dots, k-1$.

Proof. Suppose that G has n nodes. Let $\mathbf{x}$ be a nonzero integral weight vector for G. Let $G^i_{\mathbf{x}}$ denote the graph where each node v of $V(G^i)$ is replaced by a clique of x_v nodes. Recall that

$$\chi(G^i_{\mathbf{x}}) \leq \chi_f(G^i, \mathbf{x}) + |V(G^i)| \leq \chi_f(G^i, \mathbf{x}) + n; \tag{7.1}$$

see, for example, the proof of Lemma 2.1.

For each $i = 0, \dots, k-1$ consider a coloring of $G^i, \mathbf{x}$ which uses the colors $\{1, 2, \dots, \chi(G^i_{\mathbf{x}})\}$. To a node $v \in V_i$ which is colored with the x_v colors $\{c_1, \dots, c_{x_v}\}$, assign the x_v new colors $\{kc_1 - i, \dots, kc_{x_v} - i\}$. This yields a k-feasible assignment for $G, \mathbf{x}$, and so

$$\text{span}_k(G, \mathbf{x}) \leq k \max_i \{\chi(G^i_{\mathbf{x}})\} \leq k \max_i \{\chi_f(G^i, \mathbf{x})\} + kn$$

by (7.1). Hence since $\omega(G^i, \mathbf{x}) \leq \omega_k(G, \mathbf{x})$ for each i,

$$\begin{aligned}\frac{\text{span}_k(G, \mathbf{x})}{\omega_k(G, \mathbf{x})} &\leq k \max_i \left\{\frac{\chi_f(G^i, \mathbf{x})}{\omega(G^i, \mathbf{x})}\right\} + \frac{kn}{\omega_k(G, \mathbf{x})} \\ &\leq k \max_i \{\text{imp}(G^i)\} + \frac{n}{x_{\max}},\end{aligned}$$

and the result follows by (2.6). □

The first theorem in this section shows that for dense random graphs the k-imperfection ratio is asymptotically independent of k.

THEOREM 7.2. *Let k be a positive integer, and let $0 < p < 1$. Then for any $\eta > 0$, a.s.*

$$\frac{n}{4 \log_{\frac{1}{p}} n \log_{\frac{1}{q}} n} \leq \text{imp}_k(G_{n,p}) \leq (1 + \eta) \frac{n}{4 \log_{\frac{1}{p}} n \log_{\frac{1}{q}} n},$$

where $q = 1 - p$.

Proof. First we consider the lower bound. A simple first moment argument shows that the following two conditions on $G_{n,p}$ are a.s. satisfied (see, for example, [4]):

$$\alpha(G_{n,p}) \le 2\log_{\frac{1}{q}} n \qquad \text{and} \qquad \omega(G_{n,p}) \le 2\log_{\frac{1}{p}} n.$$

In addition, it is easy to verify that a.s. $\omega(G_{n,p}) \ge k$, and hence by Lemma 3.1, we have a.s.

$$\mathrm{imp}_k(G) \ge \frac{\chi_f(G_{n,p})}{\omega(G_{n,p})} \ge \frac{n}{\alpha(G_{n,p})\omega(G_{n,p})} \ge \frac{n}{4\log_{\frac{1}{p}} n \log_{\frac{1}{q}} n}.$$

Now we consider the upper bound. By Lemma 7.1, for any $t \ge 0$

$$P(\mathrm{imp}_k(G_{n,p}) > tk) \le kP(\mathrm{imp}(G_{\lceil n/k\rceil,p}) > t),$$

and hence by Theorem 3.3 of [8], we have a.s.

$$\begin{aligned}\mathrm{imp}_k(G_{n,p}) &\le k\,(1+\eta/2)\frac{\lceil n/k\rceil}{4\log_{\frac{1}{p}}\lceil n/k\rceil\log_{\frac{1}{q}}\lceil n/k\rceil}\\ &\le (1+\eta)\frac{n}{4\log_{\frac{1}{p}} n\log_{\frac{1}{q}} n} \qquad \text{for sufficiently large } n. \qquad \square\end{aligned}$$

The next result corresponds to Theorem 3.5 of [8], which shows that for suitable sparse random graphs $G_{n,p}$, $\mathrm{imp}(G_{n,p})$ is about $np/(4\ln np)$. Now we may allow slightly denser graphs and see that $\mathrm{imp}_k(G_{n,p})$ is about $np/(2k\ln np)$ when $k \ge 2$. Note that this formula depends on k, in contrast to the dense case, and it does not give the correct answer for $k = 1$.

THEOREM 7.3. *Let $k \ge 2$, and suppose that $p = p(n)$ satisfies $np \to \infty$ as $n \to \infty$ but $p = o(n^{-2/(k+1)})$. Then for any $\varepsilon > 0$, a.s.*

$$(1-\varepsilon)\frac{np}{2k\ln np} \le \mathrm{imp}_k(G_{n,p}) \le (1+\varepsilon)\frac{np}{2k\ln np}.$$

Proof. Since $p = o(1)$ and $np \to \infty$ as $n \to \infty$, for any $\varepsilon > 0$ we have a.s.

$$\chi(G_{n,p}) \le (1+\varepsilon)\frac{np}{2\ln np}; \tag{7.2}$$

see [16]. The required upper bound on the k-imperfection ratio now follows by Lemma 3.1(b).

For the lower bound, assume that $0 < \varepsilon < 1$, and let $\delta > 0$ satisfy $(1-\delta)/(1+\delta) \ge 1-\varepsilon$. By [5], a.s.

$$\alpha(G_{n,p}) \le (1+\delta)\frac{2\ln np}{p}.$$

Also, the expected number of cliques with $k+1$ nodes is $\binom{n}{k+1}p^{\binom{k+1}{2}}$, which is at most $n^{k+1}p^{\frac{k(k+1)}{2}}$. Hence the probability that the number of cliques with $k+1$ nodes in $G_{n,p}$ is at least δn is at most $(np^{\frac{k+1}{2}})^k/\delta$. Since $np^{\frac{k+1}{2}} = o(1)$, there is a.s. an induced subgraph H of $G_{n,p}$ on at least $n - \delta n$ nodes with $\omega(H) \le k$. But then a.s.

$$\chi_f(H) \ge \frac{n-\delta n}{\alpha(H)} \ge \frac{1-\delta}{1+\delta}\frac{np}{2\ln np} \ge (1-\varepsilon)\frac{np}{2\ln np}.$$

Hence by Lemma 3.1(a), a.s.

$$\operatorname{imp}_k(G_{n,p}) \geq \operatorname{imp}_k(H) \geq (1-\varepsilon)\frac{np}{2k\ln np},$$

as required. □

There is a similar result for random r-regular graphs $G_{n,r}$, which are graphs taken uniformly at random from the set of all r-regular graphs on the n nodes $\{1, 2, \ldots, n\}$ (where rn is even). The limit in the following theorem refers to $n \to \infty$ with n restricted to even integers if r is odd.

THEOREM 7.4. *Let $k \geq 2$. For each integer $r \geq 2$, there exists $\varepsilon = \varepsilon(r) > 0$ such that $\varepsilon(r) \to 0$ as $r \to \infty$ and such that for each fixed $r \geq 2$, a.s.*

$$\frac{r}{2k\ln r} \leq \operatorname{imp}_k(G_{n,r}) \leq (1+\varepsilon)\frac{r}{2k\ln r}.$$

Proof. We may argue much as in the proof of Theorem 7.3. To do this, the upper bound (7.2) has to be replaced by the following result from [6]: for each $r \geq 2$, there exists $\varepsilon = \varepsilon(r) > 0$ with $\varepsilon(r) \to 0$ as $r \to \infty$, such that

$$\chi(G_{n,r}) \leq (1+\varepsilon)\frac{r}{2\ln r}.$$

The lower bound follows from the result that $G_{n,r}$ a.s. contains a triangle-free induced subgraph H with $\chi_f(H) \geq r/2\ln r$; see the proof of Theorem 3.6 of [8]. □

REFERENCES

[1] N. Alon, Z. Tuza, and M. Voigt, *Choosability and fractional chromatic numbers*, Discrete Math., 165/166 (1997), pp. 31–38.

[2] M. Chudnovsky, N. Robertson, P. D. Seymour, and R. Thomas, *The Strong Perfect Graph Theorem*, manuscript.

[3] M. Bellare, O. Goldreich, and M. Sudan, *Free bits, PCPs and, nonapproximability—towards tight results*, SIAM J. Comput., 27 (1998), pp. 804–915.

[4] B. Bollobás, *Random Graphs*, Academic Press, London, 1985.

[5] A. M. Frieze, *On the independence numbers of random graphs*, Discrete Math., 81 (1990), pp. 171–175.

[6] A. M. Frieze and T. Łuczak, *On the independence numbers of random regular graphs*, J. Combin. Theory Ser. B, 54 (1992), pp. 123–132.

[7] S. Gerke and C. McDiarmid, *Graph imperfection*, J. Combin. Theory Ser. B, 83 (2001), pp. 58–78.

[8] S. Gerke and C. McDiarmid, *Graph imperfection* II, J. Combin. Theory Ser. B, 83 (2001), pp. 79–101.

[9] S. Gerke and C. McDiarmid, *Channel assignment with large demands*, Ann. Oper. Res., 107 (2001), pp. 143–159.

[10] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[11] A. Gräf, M. Stumpf, and G. Weissenfels, *On coloring unit disk graphs*, Algorithmica, 20 (1998), pp. 277–293.

[12] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, 2nd ed., Springer-Verlag, Berlin, 1993.

[13] W. K. Hale, *Frequency assignment: Theory and applications*, Proc. IEEE, 68 (1980), pp. 1497–1514.

[14] R. Leese and S. Hurley, eds., *Methods and Algorithms for Radio Channel Assignment*, Oxford University Press, Oxford, UK, 2003.

[15] L. Lovász, *On the ratio of optimal integral and fractional covers*, Discrete Math., 13 (1975), pp. 383–390.

[16] T. Łuczak, *The chromatic number of random graphs*, Combinatorica, 11 (1991), pp. 45–54.

[17] C. McDiarmid, *Channel assignment and discrete mathematics*, in Recent Advances in Theoretical and Applied Discrete Mathematics, C. Linhares-Salas and B. Reed, eds., Springer-Verlag, New York, 2003.

[18] C. McDiarmid and B. Reed, *Channel assignment and weighted colouring*, Networks, 36 (2000), pp. 114–117.

[19] R. McNaughton, *Scheduling with deadlines and loss functions*, Management Sci., 6 (1959), pp. 1–12.

[20] A. Schrijver, *Theory of Integer and Linear Programming*, John Wiley, New York, 1986.

[21] E. R. Scheinerman and D. H. Ullman, *Fractional Graph Theory*, John Wiley, New York, 1997.

[22] N. Schabanel, S. Ubeda, and J. Zerovnik, *A Note on Upper Bounds for the Span of the Frequency Planning in Cellular Networks*, manuscript, 1999.

[23] B. Toft, *Some problems and results related to subgraphs of colour critical graphs*, in Graphen in Forschung und Unterricht, Festschr. K. Wagner, R. Bodendiek, H. Schumacher, and G. Walther, eds., Barbara Franzbecker, Bad Salzdetfurth-Hildesheim, 1985, pp. 178–186.

[24] A. C. Tucker, *The strong perfect graph conjecture for planar graphs*, Canad. J. Math., 25 (1973), pp. 103–114.

SIAM J. Discrete Math.
Vol. 17, No. 3, pp. 426–445

ERROR EXPONENTS OF EXPANDER CODES UNDER LINEAR-COMPLEXITY DECODING*

ALEXANDER BARG† AND GILLES ZÉMOR‡

Abstract. A class of codes is said to reach capacity $\mathcal{C}$ of the binary symmetric channel if for any rate $R < \mathcal{C}$ and any $\varepsilon > 0$ there is a sufficiently large N such that codes of length $\geq N$ and rate R from this class provide error probability of decoding at most ε, under some decoding algorithm.

The study of the error probability of expander codes was initiated by Barg and Zémor in 2002 [*IEEE Trans. Inform. Theory*, 48 (2002), pp. 1725–1729], where it was shown that they attain capacity of the binary symmetric channel under a linear-time iterative decoding with error probability falling exponentially with code length N. In this work we study variations on the expander code construction and focus on the most important region of code rates, close to the channel capacity. For this region we estimate the decrease rate (the error exponent) of the error probability of decoding for randomized ensembles of codes. The resulting estimate gives a substantial improvement of previous results for expander codes and some other explicit code families.

Key words. Elias radius, error exponents, expander codes, iterative decoding

AMS subject classifications. 94B60, 05C50

DOI. 10.1137/S0895480102403799

1. Introduction.

We study transmission of information with linear codes over the *binary symmetric channel* (BSC). An $[N, K]$ binary linear code C is a K-dimensional linear subspace of $\{0, 1\}^N$. The number N is called the length of the code, and the relative dimension K/N is called the rate of the code, denoted by $R = R(C)$. A binary digit sent over the BSC is received correctly with probability $1-p$ and flipped with probability $p < 1/2$. The objective of the decoder is to restore correctly the transmitted code vector x. *Maximum likelihood* (or complete) *decoding* of C consists of choosing a codeword x' closest to the received vector y. The event that $x \neq x'$ corresponds to a decoding error. The probability $P_e(C)$ that decoding goes wrong is independent of the transmitted codeword and is a polynomial function of p. This polynomial is notoriously difficult to compute exactly, but estimating the value $P_e(C)$ can be somewhat simplified: when p is small enough, $P_e(C)$ behaves like its lowest-degree term, and the lowest degree equals half the minimum Hamming distance of C. For this reason, combinatorial coding theory is concerned with the construction of large codes with large minimum distance.

However, for both theoretical and practical reasons (like the emergence of mobile communications and their very noisy channels), there has been a renewed interest in studying the situation when p is large or, equivalently, when R is close to capacity. Shannon's theorem [11] states that $P_e(C)$ can stay close to zero only as long as we have $R < \mathcal{C}$, where $\mathcal{C}$ is the *channel capacity* and depends only on p. Furthermore, when R is close to $\mathcal{C}$, we know that for large N, for fixed rate $R < \mathcal{C}$, and for

*Received by the editors March 12, 2002; accepted for publication (in revised form) August 5, 2003; published electronically January 22, 2004.

http://www.siam.org/journals/sidma/17-3/40379.html

†DIMACS, Rutgers University, 96 Frelinghuysen Rd., Piscataway, NJ 08854 (abarg@ieee.org). This author was supported in part by Binational Science Foundation (USA-Israel) grant 1999099 and NSF grant CCR-031096.

‡École Nationale Supérieure des Télécommunications, 46 rue Barrault, 75 634 Paris 13, France (zemor@infres.enst.fr). This author's research was done in part while visiting DIMACS, Rutgers University, Piscataway, NJ, in August, 2001.

the best possible codes of rate R, the decoding error probability $P_e(C)$ takes the form $P_e(C) = 2^{-E(R,p)N+o(N)}$, where $E(R,p)$, called the *error exponent*, is a positive quantity that depends only on p and the rate R and can be computed exactly [8]. Unfortunately, the only known way to achieve this best possible error exponent is to use a random code C together with decoding algorithms that find the closest codeword by essentially performing an exhaustive search over C.

Turning to manageable decoding algorithms, until fairly recently only one class of codes was known to achieve a nonzero decoding exponent in polynomial time (though less than $E(R,p)$) for rates arbitrarily close to channel capacity: the class of concatenated codes, introduced by Forney [9] and extensively studied through the mid-1980s (see [1], [7] for overviews).

In the 1990s the discovery of turbo-codes [4] with their largely unexplained close-to-capacity performance shifted emphasis to iterative decoding techniques. One particular class of codes that can be iteratively decoded is that of expander codes. An expander code is constructed by assigning binary digits to *edges* of a bipartite graph in a way that was introduced by Tanner [16]. A surge of interest in them occurred after it was shown in [15] that if the underlying graph is an expander graph, then they correct an $\Omega(N)$ number of errors under an $O(N)$ iterative decoding. Not only was this a significant achievement for iterative decoding, but it was the first example of this kind in coding theory at large; indeed, the concatenated decoding rival requires an $O(N^2)$ decoding time.

In our work [2] we employ the iterative decoding algorithm of [17] to show that expander codes actually reach the capacity of the BSC with a positive error exponent. Though the error exponent of concatenated codes [9] is better than that of expander codes of [2], their performance again relies on a quadratic-time decoding algorithm, as opposed to a linear-time decoding algorithm in [17], [2].

An expander code C in [15], [17] is constructed from a bipartite Δ-regular graph $G = [A \cup B, E]$ with $|A| = |B| = n$ and a binary code $[\Delta, R\Delta]$ code C_0. Coordinates of a codeword in C are in one-to-one correspondence with edges of G (unlike some other constructions of codes on bipartite graphs where the coordinates of the code correspond to vertices) and satisfy the condition that the subvector incident to every vertex $v \in A \cup B$ is a code vector in C_0. The decoding algorithm of [17] consists of iterations that alternate between A and B. In each iteration all the vertices of the respective part are decoded in parallel with the code C_0. [17] also provides a bound on the number of errors correctable by this decoding.

Among the new ideas introduced in [2] is the use of two different codes: C_0 for the part A and C_1 for B. The value of the resulting error exponent is then optimized on the choice of the rates of these codes. Another idea of [2] is associating with each edge of G t binary digits of the codeword for some constant t rather than one digit in earlier constructions. (Alternatively, this can be viewed as replacing each edge by t parallel edges.) It turns out that the parameters and performance of the code C can be improved if we view the constituent codes both as binary linear codes and q-ary additive codes, $q = 2^t$. See more on this in section 4.3.

In this paper we obtain a substantial improvement of the estimate of the error exponent for expander codes, focusing on R close to capacity (Theorem 6.1 and Figure 6.1). In particular, we surpass in this region the error exponent of Forney's concatenated codes [9], a benchmark for a long time. The improvement relies on the following ideas, which were not employed in earlier analysis. In the first iteration, again relying on the q-ary structure of the code C_0, we employ detailed information on the error events. Namely, it can be shown that in the event of a decoding error the distance from the transmitted code vector to the decoded vector will most likely concentrate around some particular value. This restricts the possibilities for the incorrectly decoded code vectors of C_0. We can use this fact to make a stronger statement about the decoding error probability of the code C_1 in the second decoding iteration. Finally, we modify the original construction by adjoining a number of vertices of degree one to the expander graph. This decreases the error correcting potential but improves the overall code rate, and the resulting trade-off improves the overall error exponent. This idea borrows from turbo-codes for which the underlying graph has many degree one vertices, contrary to other classes of codes amenable to iterative decoding.

The rest of the paper is organized as follows. In sections 2 and 3 we introduce the necessary coding background. In section 4 we summarize previous work on expander codes. The new results start with section 5, where we introduce our new variation on decoding and give a refined probabilistic analysis of its behavior: this results in a first error exponent in Theorem 5.2. Already, this result improves the best previously known error bound for expander codes [2]. In section 6 this is improved to Theorem 6.1 through a modified construction that we analyze. Finally, we give some concluding comments.

2. Codes and their parameters. In this section we introduce basic notation and recall some bounds on the parameters of codes used in deriving properties of the expander code construction. Although our ultimate goal will be binary codes, we will also consider codes over larger alphabets of size $q = 2^t$. By $\mathcal{H}_q = \mathcal{H}_q^N$ we denote the q-ary Hamming space, i.e., the N-dimensional coordinate space over the field of q elements. The number of nonzero coordinates of a vector $x \in \mathcal{H}_q$ is called the (Hamming) weight, denoted $|x|$. The Hamming distance is defined by $\mathrm{d}(x, y) = |x-y|$.

For a given linear code C the minimum weight of a nonzero codeword in C is called its *distance.* For a given q we will use notation $C[N, K, D]$ to refer to a linear code of length N, dimension K, and distance $D = D(C)$, occasionally omitting the distance.

One of the key problems of combinatorial coding theory is finding the maximum size of a code C of length N and distance D. Consider families of codes $C_i, i = 1, 2, \dots,$ of growing length N_i, rate R_i, and relative distance $\delta_i = D(C_i)/N_i$. According to the well-known Gilbert–Varshamov (GV) bound there exist sequences of codes of rate $R(C_i) \to R$ and relative distance $\delta_i \to \delta$ for any

$$R < 1 - H_q(\delta),$$

where

$$H_q(x) = -x \log_q \frac{x}{q-1} - (1-x)\log_q(1-x)$$

is the q-ary entropy function. For a given R we let $\delta_{\mathrm{GV}}^{(q)}(R)$ denote the GV distance: $\delta_{\mathrm{GV}}^{(q)}(R) = H_q^{-1}(1-R)$. Note that $\delta_{\mathrm{GV}}^{(q)}(0) = (q-1)/q$, $\delta_{\mathrm{GV}}^{(q)}(1) = 0$. For $q = 2$ we omit the superscript and write simply $\delta_{\mathrm{GV}}(R)$. Likewise, throughout the paper, if the base of the logarithms and exponents is missing, it is equal to 2.

A number of upper bounds are known on the relative distance $\delta(R)$ of a code sequence of rate R. We mention the Bassalygo–Elias bound which asserts that, for binary codes

$$\delta(R) \le \delta_E(R) := 2\delta_{\mathrm{GV}}(R)(1 - \delta_{\mathrm{GV}}(R)).$$

We call the quantity $\delta_E(R) := 2\delta_{\mathrm{GV}}(R)(1-\delta_{\mathrm{GV}}(R))$ the *Elias radius*; its relevance to decoding will become apparent from Proposition 3.1 below.

As $q = 2^t$ gets large, the quantity $(q-1)/q$ comes close to one, and the inverse entropy function gets close to the linear function $1-R$. More precisely, for any $\beta > 0$ there exists a value t_0 such that for all $t > t_0$,

$$0 < (1-R) - \delta_{\mathrm{GV}}^{(q)}(R) < \beta. \tag{2.1}$$

3. Decoding and error exponents. This section is devoted to some properties of random linear codes related to their maximum likelihood decoding, which will be used in the analysis of expander decoding.

3.1. Decoding. Let C be an $[N,K]$ binary code used on a BSC(p), and let $P_e(C)$ be its average probability of decoding error under maximum likelihood decoding. By the classical results of coding theory [8], [11] there exist sequences of binary linear codes such that the probability $P_e(C)$ under maximum likelihood decoding falls exponentially with the code length N. Therefore, define the error exponent $E(C) = -N^{-1}\log P_e(C)$. We define the best attainable error exponent for the rate R as

$$E(R,p) = \liminf_{N\to\infty} \max_{C\subset\mathcal{H}_2^N:\, R(C)\ge R} E(C).$$

It was proved in [8] (see also [11]) that $E(R) > 0$ for $0 \le R < \mathcal{C}$.

3.2. Random coding exponent and typical events. The best known existence (lower) bound $E_0(R,p)$ on the error exponent $E(R,p)$ of binary codes (linear or not) is obtained by the random coding method [11]. The function $E_0(R,p)$ is positive for all rates below the channel capacity $\mathcal{C} = 1 - H(p)$. It is easy to prove by random choice [11] that there exist sequences of binary linear $[N, RN]$ codes that attain the GV bound on the minimum distance and reach the error exponent $E_0(R,p)$ under maximum likelihood decoding. We assume that p is fixed and R varies from zero to $\mathcal{C}$. The form of the bound depends on the location of R with respect to $\mathcal{C}$. We are interested in the high-rate region, i.e., R close to $\mathcal{C}$. For high rates we obtain the bound [8], [11]

$$E_0(R,p) = D(\delta_{GV}(R)\|p),$$

where

$$D(x\|y) := x\log(x/y) + (1-x)\log((1-x)/(1-y)).$$

This bound is actually tight in the region $R_{\mathrm{crit}} \le R \le \mathcal{C}$, where the value $R_{\mathrm{crit}} = 1 - H(\rho_0)$,

$$\rho_0 = \frac{\sqrt{p}}{\sqrt{p}+\sqrt{1-p}}, \tag{3.1}$$

is called the critical rate of the channel. In other words, for $R_{\mathrm{crit}} \le R \le \mathcal{C}$ we have $E(R,p) = E_0(R,p)$ (see [8]). Note that $E_0(\mathcal{C},p) = 0$.

For rates $0 \le R \le R_{\mathrm{crit}}$ the random coding exponent has the following form:

$$E_0(R,p) = -\delta_{\mathrm{GV}}(R)\log 2\sqrt{p(1-p)} \quad (0 \le R \le R_x),$$
$$E_0(R,p) = D(\rho_0\|p) + R_{\mathrm{crit}} - R \quad (R_x \le R \le R_{\mathrm{crit}}),$$

where $R_x = 1 - H(2\rho_0(1-\rho_0))$.

Both construction and decoding complexity of known code sequences that attain this error exponent grow exponentially with the code length N.

It is possible to say more about the geometry of typical error events under maximum likelihood decoding of random codes. Namely, in the high-rate region which interests us, conditional on the event of a decoding error, the typical relative distance between the transmitted codeword and the decoded codeword is close to $\delta_E(R)$. Formally we have the following proposition.

PROPOSITION 3.1. *Let p be the parameter of a BSC. Assume that the transmitted vector is the all-zero one, and denote by z the decoded vector output by maximum likelihood decoding (this is a random variable which depends on the noise realization).*

For any R and for any large enough N, there exists a code C such that, for any p such that $R_{crit} \le R < \mathcal{C}$,

- *C has error exponent $E_0(R,p)$;*
- *for any $\alpha > 0$ we have*

$$\Pr\left[\left|\delta_E(R) - \frac{|z|}{N}\right| \ge \alpha \quad \middle| \quad z \ne 0\right] < 2^{-Nc(\alpha)} \quad (c(\alpha) > 0 \textit{ is independent of } N),$$

where $\Pr[\cdot]$ *is the probability that the random vector z fulfills the condition in the brackets.*

The proof of this result (see the appendix) relies on a combination of facts known to coding theorists but not spelled out in the literature.

3.3. Error exponents of concatenated codes. An important example of codes that attain channel capacity under low-complexity decoding is given by Forney's concatenated codes [9] (see also [7]). Concatenated codes form a generalization of Elias's product codes and provide code families with better parameters in terms of both code distance and error exponents.

Binary $[N = nm, k\ell]$ concatenated codes are constructed by first encoding the k q-ary message symbols, $q = 2^\ell$, with an $[n,k]$ q-ary code C_1, then representing every code symbol back as a string of ℓ bits, and then encoding it with an $[m,\ell]$ binary code C_0. We assume that both rates $R_0 = \ell/m$ of the inner code C_0 and $R_1 = k/n$ of the outer code are fixed.

Let $m = \log_2 N$. As with all linear codes, properties of a typical concatenated code found by random choice are much better than those of the known explicit families. In particular, there exists a sequence of binary $[m, \ell = mR_0]$ codes C_0 for which the error probability of maximum likelihood decoding falls as $2^{-mE_0(R_0,p)}$. Moreover, a brute-force implementation of decoding of the code C_0 has complexity $O(m2^{R_0 m}) = O(N \log N)$. An explicit family is obtained by taking as outer codes a sequence of Reed–Solomon codes C_1 of growing length n over the alphabet of growing size $q = 2^\ell$. Performing algebraic (generalized minimum distance) decoding of the code C_1 with complexity $O(n^2)$, we obtain the error exponent (see [9])

$$E_{\mathrm{F}}(R) = \max_{R<R_0<\mathcal{C}} E_0(R_0,p)\left(1 - \frac{R}{R_0}\right).$$

The overall decoding complexity is $O(N^2)$.

The error exponent E_{F} was improved by Blokh and Zyablov [5] (see also [7]), who consider *multilevel* concatenated codes: this involves replacing the inner code C_0 by

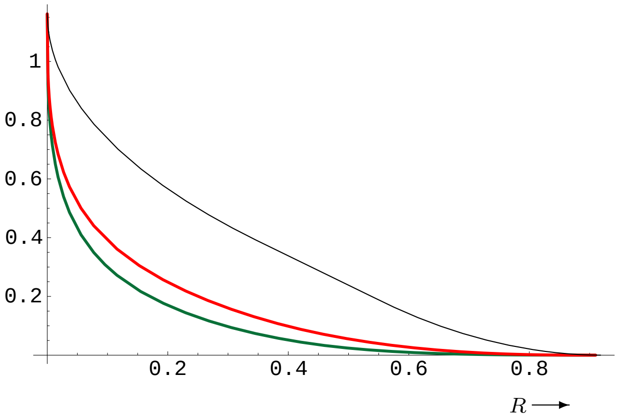


FIG. 3.1. *Error exponents of concatenated codes;* $p = 0.01$*,* $\mathcal{C} \approx 0.919$*. Bottom to top: the exponent* E_F *of concatenated codes; bound* (3.2) *with* $i = 5$*; the random coding bound* $E_0(R, p)$*.*

i different binary codes and the outer code C_1 by i Reed–Solomon codes of different rates. They obtain the exponent

$$E_i(R) = \max_{R<R_0<\mathcal{C}} \left(\frac{i(R_0 - R)}{R_0 \sum_{j=1}^{i} \left(E_0(R_0 \frac{i-j+1}{i}, p)\right)^{-1}} \right) \qquad (3.2)$$

under an $O(N^2)$ multistage decoding algorithm. We have $E_{\mathrm{F}}(R) = E_1(R)$, and

$$E_i(R) < E_{i+1}(R), \quad 0 < R < \mathcal{C}, \ i = 1, 2, \dots;$$

see Figure 3.1.

We note that for code rates not too close to capacity, the quotient $E_0(R,p)/E_{\mathrm{F}}(R)$ is uniformly bounded from above by a function that depends only on p. Moreover, for $R \to 0$ we even obtain $E_{\mathrm{F}}(R) \sim E_0(R,p)$. The situation changes dramatically for $R \to \mathcal{C}$. Indeed, letting $\mathcal{C} - R = \varepsilon$, we obtain $E_0(R,p) \approx c(p)\varepsilon^2$, where $c(p)$ depends only on the channel. In contrast, $E_{\mathrm{F}}(R) = O(\varepsilon^3)$. Thus, the quotient $E_0(R,p)/E_{\mathrm{F}}(R) \to \infty$ as $\varepsilon \to 0$. This is also the case for the multilevel exponents: we have $\lim_{R\to\mathcal{C}}(E_0(R,p)/E_i(R)) = \infty$ for any finite i and even in the limit $i \to \infty$. Improving error exponents of low-complexity decoding algorithms in the range of code rates R close to $\mathcal{C}$ is therefore of most interest. In this paper we improve the constant factor in front of the ε^3 compared to E_{F}, while the problem of improving the decrease order from ε^3 remains an open question.

4. Expander graphs and expander codes.

4.1. Expander graphs. Consider a balanced Δ-regular bipartite graph $G = [A \cup B, E]$ with the vertex set $A \cup B$, where $|A| = |B| = n$ and where every edge has one endpoint in A and one in B. Let λ be the second eigenvalue of G. We have $|E| = N := n\Delta$. Any vertex $v \in A$ $(v \in B)$ will be called a left (right) vertex.

For a vertex v and a subset of vertices S, denote by $d_S(v)$ the S-degree of v, i.e., the number of edges that connect v to some vertex of S.

We need one result from [17].

Lemma 4.1 (see [17]). *Let $S \subset A$ and $T \subset B$. Then the average degree $\overline{d}_{ST}$ of the subgraph induced by $S \cup T$ satisfies*

$$\overline{d}_{ST} \le \frac{2|S||T|}{|S|+|T|}\frac{\Delta}{n} + \lambda.$$

This lemma implies that expander graphs approximate well the behavior of random bipartite graphs. Namely, if λ is small compared to Δ, the size of the set T of right vertices whose S-degrees exceed substantially the expected S-degree of a vertex in a random graph can be made arbitrarily small by choosing a sufficiently large Δ. This is made formal in the following lemma.

Lemma 4.2. *Let $S \subset A$ with $|S| = s = \sigma n$. Let $\alpha > \alpha_\sigma$, where $\alpha_\sigma = \frac{\lambda}{2\sigma\Delta}$. Let T be the subset of B defined by $T = \{v \in B,\ d_S(v) \ge (1+\alpha)\sigma\Delta\}$. Then we have*

$$|T| \le \frac{\alpha_\sigma}{\alpha - \alpha_\sigma}|S|.$$

Proof. The number of edges of the graph $G_{S\cup T}$ induced by $S \cup T$ is at least $|T|(1+\alpha)\sigma\Delta$, and therefore the average degree $\overline{d}_{ST}$ satisfies

$$\frac{2|T|(1+\alpha)\sigma\Delta}{|S|+|T|}.$$

Applying Lemma 4.1, we therefore have

$$\frac{2|T|(1+\alpha)\sigma\Delta}{|S|+|T|} \le \frac{2|S||T|}{|S|+|T|}\frac{\Delta}{n} + \lambda,$$

whence

$$2|T|(1+\alpha)\sigma\Delta \le 2|S||T|\frac{\Delta}{n} + \lambda(|S|+|T|),$$

$$|T|\left(\alpha\sigma\Delta - \frac{\lambda}{2}\right) \le \frac{\lambda}{2}|S|,$$

and the result follows after some rearranging. □

4.2. Expander codes. Let $G = [A \cup B, E]$ be as above, with $\lambda \ll \Delta$. Let us fix an arbitrary order of the edges in E. For a vertex v this defines an ordering of edges $v(1), \dots, v(\delta)$ incident to it. Given a binary vector $x \in \mathcal{H}_2^N$, this ordering induces a subvector $x_v = (x_{v(1)}, \dots, x_{v(\Delta)})$.

To construct a linear code C associated with the graph G we also need two binary codes, $C_0[\Delta, R_0\Delta, \delta_0\Delta]$ and $C_1[\Delta, R_1\Delta, \delta_1\Delta]$. An N-vector x is a codeword of C if and only if for every left vertex v the subvector x_v is a codeword in C_0 and for every right vertex w the subvector x_w is a codeword in C_1. The code C has length $N = n\Delta$ and rate $R \ge R_0 + R_1 - 1$.

Codes associated with bipartite graphs were considered in [2], [6], [10], [14], [15], [16], [17]. In particular, Sipser and Spielman [15] introduced an important idea of estimating properties of a simple iterative decoding procedure via spectral properties of the graph G. The main result of [15] is given by the following theorem.

Theorem 4.3 (see [15]). *For any $\varepsilon > 0$ there exists a polynomial-time constructible code with relative distance $\delta - \varepsilon$ and rate $1 - 2H(\sqrt{\delta})$ for which any $\alpha < \delta/48$ fraction of errors can be corrected by a circuit of size $O(N \log N)$ and depth $O(\log N)$. The complexity of a sequential implementation of this decoding is $O(N)$.*

This theorem gives codes with positive rate $R > 0$ for $0 < \delta < 0.011$. We note that [15] used the above family of expander codes with $C_0 = C_1$ but with a somewhat different decoding algorithm than the ones used in this paper. The present construction together with the decoding procedure (5.1) below was used in [17] to improve the fraction of correctable errors in Theorem 4.3 to $\delta/4$.

4.3. Replicated expander codes. A modification of the above construction was introduced in [2]. Namely, assume that every edge $(v, w) \in E$ is really a bundle of t parallel edges, each with one end in v and the other in w. Further, assume that C_0 is a $[t\Delta, t\Delta R_0]$ code and C_1 a $[t\Delta, t\Delta R_1]$ code. Both codes can be considered as binary linear codes and also as q-ary additive codes (additive subgroups of $\mathbb{F}_q, q = 2^t$).

The purpose of introducing replication is to make use of (2.1). In particular, with $C_0 = C_1$ we obtain the following theorem.

THEOREM 4.4 (see [2]). *For any $R_0, 0 < R_0 < 1$, and $\varepsilon > 0$ there exists an expander code C of rate $R \geq 2R_0 - 1$ and relative distance $\delta = (1-R_0)H^{-1}(1-R_0)-\varepsilon$. Iterative decoding applied to this code corrects any $\alpha < \delta/4$ fraction of errors.*

This is an improvement over Theorem 4.3: in particular, we obtain codes C with positive rates for all $0 \leq \delta < 0.055$. This is the best result known to date for the fraction of errors correctable in linear time with expander codes.

5. Attaining capacity with linear complexity.

5.1. A simple bound. Replication also enables us to obtain good estimates of the error exponent of iterative decoding. Consider the following decoding algorithm of an expander code C. For a vector $y \in \mathcal{H}_2^N$ let $L(y)$ $(R(y))$ be the vector z such that for every $v \in A$ $(v \in B)$ the vector z_v is one of the codewords of C_0 (C_1) closest to y_v. Suppose that the transmitted zero vector is received as $y \neq 0$. Consider the decoding procedure of [17], [2]:

$$y^{(0)} = y,\ y^{(1)} = L(y^{(0)}),\ y^{(2)} = R(y^{(1)}),\ y^{(3)} = L(y^{(2)}) \ldots. \tag{5.1}$$

We assume that in computing $L(y^{(0)})$ the algorithm relies on the representation of C_0 as a binary code. In other words, for every left vertex v the associated binary subvector $(y^{(0)}_{v(1)}, y^{(0)}_{v(2)}, \ldots, y^{(0)}_{v(t\Delta)})$ is decoded with the binary code $C_0[t\Delta, t\Delta R_0]$ into one of the closest code vectors of C_0. All the subsequent stages use the q-ary structure of C_1 and C_0. For instance, in the second stage this amounts to grouping consecutive groups of t bits of the vector $y^{(1)}$ into q-ary symbols. More precisely, for every right vertex w the q-ary subvector $y^{(1)}_w = (y^{(1)}_{w,1}, y^{(1)}_{w,2}, \ldots, y^{(1)}_{w,\Delta})$ associated with it can be written in binary representation as

$$[(y^{(1)}_{w(1)}, \ldots, y^{(1)}_{w(t)}), \ldots, (y^{(1)}_{w((i-1)t+1)}, \ldots, y^{(1)}_{w(it)}), \ldots, (y^{(1)}_{w((\Delta-1)t+1)}, \ldots, y^{(1)}_{w(\Delta t)})].$$

All the n Δ-subvectors y_w are independently decoded with the q-ary code C_1 according to the minimum of the q-ary Hamming distance.

The procedure stops after either having met a fixed point (i.e., when $y^{(i+2)} = y^{(i+1)} = y^{(i)}$ for some i) or after having made $O(\log N)$ steps.

Error exponents of this decoding algorithm for expander codes (replicated or not) were analyzed in [2]. The strongest result obtained in that paper is as follows.

THEOREM 5.1 (see [2]). *For a given rate R, any $\varepsilon > 0$, and $\alpha < 1$ there exists a polynomial-time constructible family of replicated expander codes of length N such that $P_e(C, p) \leq 2^{-\alpha N F_1(R,p)}$, where*

$$F_1(R, p) = \max_{R<R_0<\mathcal{C}} E_0(R_0, p)(R_0 - R)/2 - \varepsilon. \tag{5.2}$$

The decoding complexity of these codes is the same as in Theorem 4.3.

The idea behind the proof is as follows. We take a code C_0 of rate $R_0 > R$ and a code C_1 of rate R_1 close to one. The first decoding step, $y^{(1)} = L(y^{(0)})$, removes most of the errors from the received word, which is possible because the code C_0 has relatively large distance. By taking a large but fixed Δ we ensure that, for every left vertex, the error exponent after the first step is close to the random coding bound $E_0(R_0, p)$ (section 3.2).

At the second step, most vertices correspond to subvectors containing less than $D_1/2 \approx (R_0 - R)/2$ errors and are therefore corrected. The expanding properties of the graph G ensure that the remaining atypical subvectors, call them "badly in error," are corrected at later iterations if their number is small enough. Small enough means smaller than $n\delta_1^q/2$, where $\delta_1^q = D_1/\Delta$ is the q-ary relative minimum distance of C_1. Specifically, Lemma 4.2 ensures that if at some point of the decoding procedure the number of wrongly decoded subvectors (or vertices of the graph) is less than $n\delta_1^q/2$, then the number of subvectors in error must decrease geometrically at the next iteration.

Summarizing, we simply upperbound the probability of a decoding error by the probability that the number of left subvectors remaining in error after the first iteration is more than $n\delta_1^q/2$.

Remark. For any fixed rate R the decoding complexity of expander codes is bounded above as $O(N)$. The value of the multiplicative constant depends on the code rate and increases as R approaches capacity $\mathcal{C}$. In the neighborhood of capacity this constant becomes large: for $R = \mathcal{C}(1-\gamma)$ it can be estimated to be $\exp(\gamma^{-2})$. This remark applies to Theorem 5.1 and to the subsequent constructions of this paper. Note that such a behavior was already the case for the concatenated code constructions of section 3.3.

5.2. Refined technique: Overview. What precludes one from obtaining a better error exponent in this way is the fact that at the second step the proof relies on a very strong convergence condition, namely, that most right vertices receive fewer than $D_1/2$ errors. Since the length of the code C_1 is a fixed constant, we could in principle involve more powerful decoding, but it is unclear how to bound its error rate. Moreover, incorrect codewords obtained in left vertices after the first iteration tend to be of one and the same Hamming weight (by Proposition 3.1). This information is also unclaimed in the proof of Theorem 5.1.

Our strategy will therefore be to study the probability that the number of vertices in error after the *second* iteration is sufficiently small for Lemma 4.2 to apply and again guarantee convergence of the decoding algorithm.

Our analysis is first based on the observation that, given the number σn of left vertices that are wrongly decoded by the first iteration, we know that the typical resulting error vector $y^{(1)}$ satisfies the following:

1. For almost every right vertex w, the right subvector $y_w^{(1)}$ has q-ary weight very close to $\sigma\Delta$. This is an application of Lemma 4.2.
2. Almost every edge corresponds to a binary t-tuple which is either all-zero or of weight very close to $\delta_E(R_0)t$. This is an application of Proposition 3.1.

Given this information on the right subvectors $y_w^{(1)}$, we want to use a modified maximum likelihood decoder for the code C_1 and estimate the probability that an error will occur at the second decoding iteration. However, the difficulty we face is that even if we have a typical error pattern for almost every $y_w^{(1)}$, we have little control over its probability distribution. To tackle this problem we will introduce partial randomization of the overall code construction and determine an error distribution for almost every right subvector $y_w^{(1)}$.

5.3. Refined technique: Details. We modify somewhat the code family and the decoding algorithm.

The general code construction is the same as in section 4.3. For a more detailed analysis of decoding we need some further properties of the graph G and the codes C_0, C_1. Since we will use the full power of Lemma 4.2, it is important that we choose bipartite graphs G with the smallest possible λ, e.g., Ramanujan graphs [13] with $\lambda = O(\sqrt{\Delta})$. Next, we take the code C_0 to satisfy the condition of Proposition 3.1. The code C_1 is chosen to satisfy a restriction on the nature of decoding error events made precise in Lemma 5.4 below. Finally, we introduce a small amount of randomization: namely, we consider an ensemble of expander codes obtained by choosing randomly and independently the coordinate ordering of every vertex subcode.

Two relative minimum distances are used for C_0: the *binary* relative distance $\delta_0 = d_0/(t\Delta)$, where d_0 is the binary minimum distance of C_0, and the *q-ary* relative distance $\delta_0^q = d_0^q/\Delta$, where d_0^q is the q-ary minimum distance of C_0. By choosing Δ and t large enough (but fixed) we ensure that both distances are close to the GV bounds, $\delta_0 \approx \delta_{\mathrm{GV}}(R_0)$ and $\delta_0^q \approx \delta_{\mathrm{GV}}^{(q)}(R_0)$, and that the error probability of max-likelihood decoding of C_0 behaves as

$$P_e(C_0) \approx 2^{-t\Delta E_0(R_0,p)} \quad (R \to \mathcal{C}).$$

For the code C_1 we consider only its q-ary distance $D_1 = \delta_1^q \Delta$, of which we assume $\delta_1^q \approx \delta_{GV}^{(q)}(R_1)$.

The decoding algorithm consists of alternating left and right decodings. The first (left) decoding, $L(y^{(0)})$, is the same as in (5.1). The second (right) decoding uses a modified "max-likelihood" decoding adapted to a nonsymmetric additive q-ary channel, which we describe in section 5.4. Every subsequent decoding step uses standard decoding for the q-ary symmetric channel, described in section 5.1.

The claim about the properties of this construction, proved in the remainder of section 5, is the following theorem.

THEOREM 5.2. *For a given rate R, there exists a polynomial-time constructible family of replicated expander codes of length N, defined up to the orderings of the coordinates of the constituent codes, such that, given a random set ω of orderings of the constituent codes,*

$$P_e(C, p, \omega) \le 2^{-NF(R,p)(1-\varepsilon(R,p))},$$

where $\varepsilon(R,p)$ is a function depending only on R and p such that $\varepsilon(R,p) < 1$ when $R < \mathcal{C}$ and $\varepsilon(R,p) \to 0$ when $R \to \mathcal{C}$ and where

$$F(R,p) = \max_{R<R_0<\mathcal{C}} E_0(R_0,p)\frac{R_0 - R}{H(\delta_{\mathrm{E}}(R_0))}.$$

The decoding complexity of these codes is the same as in Theorem 4.3.

Proof. Let γ_0 be the typical fraction of bits in error after decoding a corrupted codeword of C_0, given that a wrong codeword is output. By Proposition 3.1 for $R_0 \ge R_{\mathrm{crit}}$ we have $\gamma_0 = \delta_E(R_0)$. Let us denote by S the set of vertices of A in error after the first iteration and by T the set of vertices of B in error after the second iteration. Note that

$$\Pr[|S| = \sigma n] \lesssim \binom{n}{\sigma n} 2^{-t\Delta\sigma n E_0(R_0,p)} \lesssim \exp[-N\sigma E_0(R_0,p)]$$

since by choosing t and Δ sufficiently large the term $N^{-1}\log\binom{n}{\sigma n}$ can be made arbitrarily small. Here and henceforth the probability $\Pr[\cdot]$ is computed with respect to the random vector z received from the channel.

Lemma 4.2 implies that the event $|T| < \delta_0^q n/4$ is sufficient to ensure convergence of the decoding algorithm. This is because the number of vertices in error at the next iteration will have an order of magnitude that can be made as small as $(\text{const}\cdot|T|/\sqrt{\Delta})$, which we can make smaller than $\delta_1^q n/2$ by choosing a sufficiently large Δ. Under this condition the convergence of the algorithm after $O(\log n)$ iterations follows by [2, Prop. 2].

The error probability P_e under the iterative decoding algorithm is upperbounded by the probability that the convergence conditions do not hold. Let $\mathcal{F}$ be the set of multiple edges (or q-ary symbols) that are "heavy" after the first decoding iteration, i.e., have binary weight at least $(1+\alpha)^2\gamma_0 t$. We can therefore claim that for any ε, $\alpha > 0$,

$$P_e \le \sum_{s\ge D_1/2} [P_0(\varepsilon,\alpha,s) + P_1(\varepsilon,\alpha,s)], \tag{5.3}$$

where

$$P_1(\varepsilon,\alpha,s) = \Pr\Big[|S| = s, |\mathcal{F}| > \varepsilon s\Delta\Big],$$
$$P_0(\varepsilon,\alpha,s) = \Pr[A_{s,\varepsilon}]\Pr\Big[|T| \ge \frac{\delta_0^q}{2}n \mid A_{s,\varepsilon}\Big]$$

and where $A_{s,\varepsilon}$ stands for the event $A_{s,\varepsilon} = \{|S| = s, |\mathcal{F}| \le \varepsilon s\Delta\}$.

The rest of the proof is upperbounding P_e. We first obtain a rough estimate of P_1. Denote $\sigma = s/n$. An edge may be "heavy" for two reasons. It may be incident to the set $S_{\mathcal{F}} \subset S$ made up of those vertices for which left decoding of the first iteration outputs a (wrong) codeword of C_0 of binary weight at least $(1+\alpha)\gamma_0 t\Delta$. Being incident to $S_{\mathcal{F}}$ is a rare event, but given that, a high proportion of its edges are likely to be heavy. Or it may be incident to $S \setminus S_{\mathcal{F}}$, but then being heavy is in itself a rare event. We write

$$P_1 \le P_2 + P_3,$$

where

$$P_2 = \Pr\Big[|S_{\mathcal{F}}| > \frac{\varepsilon\sigma n}{2} \mid |S| = \sigma n\Big],$$
$$P_3 = \Pr\Big[|\mathcal{F}| \ge \varepsilon\sigma n\Delta \mid |S_{\mathcal{F}}| \le \frac{\varepsilon\sigma n}{2}, |S| = \sigma n\Big]. \qquad \square$$

LEMMA 5.3. *We have $P_2 \le \Pr[|S| = s]2^{-f_2(\delta_0,\alpha)\varepsilon\sigma N}$ and $P_3 \le \Pr[|S| = s]2^{-f_3(\delta_0,\alpha)\varepsilon\sigma N}$, where $f_2(\delta_0,\alpha)$ and $f_3(\delta_0,\alpha)$, both positive, stay bounded away from zero as $R \to \mathcal{C}$.*

Proof. We have

$$P_2 \le \Pr[|S| = s]\binom{\sigma n}{\varepsilon\sigma n/2}\Pr\Big[\Big|\gamma_0 - \frac{|y|}{t\Delta}\Big| \ge \alpha\gamma_0\Big]^{\varepsilon\sigma n/2}.$$

By Proposition 3.1 we can write $-\log \Pr[\cdot]^{\varepsilon\sigma n/2} \geq c(\alpha\gamma_0)\varepsilon\sigma N/2$. Compared to this, the binomial coefficient $\binom{\sigma n}{\varepsilon\sigma n/2}$ can be made to have a negligible exponent by choosing $t\Delta$ large enough. This gives an exponent $f_2(\delta_0, \alpha) \approx c(\alpha\gamma_0)/2$.

Let us now evaluate P_3. Take any vertex v of $S \setminus S_{\mathcal{F}}$, and let γ be the proportion of bits in error of the corresponding output codeword of the code C_0. By definition of $S_{\mathcal{F}}$ we have $\gamma \leq (1+\alpha)\gamma_0$. Consider any multiple edge: the random grouping of the $t\Delta$ coordinate positions of C_0 into Δ multiple edges will give $\binom{t\Delta}{t}$ possible choices for the t edges that will form the multiple edge. The probability that the multiple edge contains more than $(1+\alpha)\gamma t$ bits in error is therefore

$$\pi = \sum_{t \geq k > (1+\alpha)\gamma t} \frac{\binom{\gamma t\Delta}{k}\binom{t\Delta - \gamma t\Delta}{t-k}}{\binom{t\Delta}{t}}. \tag{5.4}$$

The dominating term in this last sum occurs for the smallest k; after rearranging and neglecting terms nonexponential in t it is fairly routine to obtain that

$$\pi \approx \gamma^{(1+\alpha)\gamma t}(1-\gamma)^{(1-(1+\alpha)\gamma)t}\binom{t}{(1+\alpha)\gamma t},$$

which gives, for small α,

$$\pi \approx e^{-\frac{\gamma}{2(1-\gamma)}\alpha^2 t},$$

so that for fixed α and γ we obtain that π is exponentially small in t in a way that does not depend on Δ.

Next observe that if we compute the probability π' that a given multiple edge has weight larger than $(1+\alpha)\gamma$, given that i other given edges have weight larger than $(1+\alpha)\gamma$, then we will obtain the formula (5.4) with Δ replaced by $\Delta' = \Delta - i$ and γ replaced by $\gamma' < \gamma$. We will therefore have $\pi' \leq \pi$. With this in mind we write that

$$P_3 \leq \Pr[|S| = s]\binom{\sigma n\Delta}{\varepsilon\sigma n\Delta/2}\pi^{\varepsilon\sigma n\Delta/2}.$$

Again, by choosing t large enough, we obtain that $\binom{\sigma n\Delta}{\varepsilon\sigma n\Delta/2}$ has negligible exponent and obtain the desired upper bound for P_3 with

$$f_3(\delta_0, \alpha) \approx \frac{1}{4\ln 2}\frac{\gamma_0}{1-\gamma_0}\alpha^2$$

for small α. □

We see that the exponents of P_2 and P_3, however small, stay larger than a constant times σN as $R \to \mathcal{C}$. It will therefore be apparent that when $R \to \mathcal{C}$, the probabilities P_2 and P_3, and hence P_1, are negligible compared to P_0. Next, we upperbound P_0 in (5.3) by writing

$$P_0 \leq \Pr[|S| = \sigma n]\Pr\left[|T| \geq \frac{\delta_0^q}{2}n \;\middle|\; |S| = \sigma n, |S_{\mathcal{F}}| \leq \varepsilon\sigma n\right].$$

Assume that $P_0 = 2^{-NE(R)}$. Since we are interested only in the exponential behavior of the sum in (5.3), we simply need to find the term P_0 with the smallest exponent. Therefore, letting

$$\Pr\left[|T| \geq \frac{\delta_0^q}{2}n \;\middle|\; A_{s,\varepsilon}\right] = 2^{-NE_T}$$

and switching to exponents, we obtain

$$E(R) \geq \min_{\sigma \geq \delta_1^q/2} (E_0(R_0, p)\sigma + E_T). \tag{5.5}$$

Next we evaluate the exponent E_T: this is where the random choice of coordinate orderings comes in. Because the coordinate orderings of the right subcodes have been chosen randomly and are independent of the error vector (they might very well have been chosen *after* the first decoding step), we can argue that each right subcode is, independently of the others, submitted to an error vector chosen equiprobably among a set of error vectors with a given weight pattern. We therefore need to estimate the typical weight pattern of the error vector on a right vertex after the first decoding step. By Lemma 4.2, all but a negligible fraction of right vertices have more than $(1+\alpha)\sigma\Delta$ edges incident to S (recall that α_σ in Lemma 4.2 can be made as small as we want because $\lambda = O(\sqrt{\Delta})$ and Δ can be taken to be arbitrarily large). This means that almost every right vertex w has an error vector of weight at most $(1+\alpha)\sigma\Delta$. Let us establish a further property of these error vectors.

Let Q' be the subset of Q defined by the set of those q-ary symbols whose binary representation is of weight at most $(1+\alpha)^2\gamma_0 t$. We now argue that almost all of the subvectors $y_w^{(1)}$ have most of their symbols in the subset Q'.

Indeed, $\mathcal{F}$ is exactly the set of symbols that do not belong to Q' after the first decoding iteration. By Lemma 5.3 we can assume that $|\mathcal{F}| \leq \varepsilon\sigma n\Delta$ because the opposite event occurs with an exponent that would be much greater than the others (and thus negligible probability) when $R \to \mathcal{C}$.

By the Markov inequality, the number of right vertices w that have more than $\beta\sigma\Delta$ incident edges that correspond to a symbol in $Q \setminus Q'$ is less than $\varepsilon n/\beta$. Now we can simultaneously choose β and ε such that β, ε, and $\lambda = \varepsilon/\beta$ are all small. Summarizing, we obtain that almost all vertices of B (i.e., $|B|(1-\eta)$ of them) have an error vector of weight at most $(1+\alpha)\sigma\Delta$ and such that all but $\beta\sigma\Delta$ of its nonzero symbols belong to Q', where β is again arbitrarily small. We now claim that

$$E_T \geq (\delta_0^q/4 - \eta)E_1, \tag{5.6}$$

where E_1 is an error exponent for the right decoder, given that the error vector has the above pattern. To see this, assume the worst, namely, that all $\eta|B|$ vertices with their error vector of the wrong pattern will be wrongly decoded. The claim now consists of saying that if not more than $n\delta_0^q/4$ right vertices are in error, then the subsequent decoding steps must converge correctly. This in turn follows from Lemma 4.2, which implies that the number of left vertices that have more than $\Delta\delta_0^q/2$ edges incident to T (the only ones that can be wrongly decoded at the third iteration) can, by choosing λ/Δ small enough, be made sufficiently small (smaller than $n\delta_1^q/4$, for example) so that the number of vertices in error will shrink geometrically at each iteration as in [17], [2], or section 5.1. The choice of the fraction 1/4 is arbitrary and can be replaced by any number less than 1/2.

We next evaluate E_1.

5.4. Decoding C_1. The right decoder assumes that the q-ary error vector has weight not more than $\sigma'\Delta$ with $\sigma' = \sigma(1+\alpha)$ and that among its nonzero symbols, not more than $\beta\sigma\Delta$ do not belong to the subset $Q' \subset Q$ (q-ary vectors with restricted binary weight of symbols). If it does not find any codeword that fits this hypothesis it returns an arbitrary codeword (say, a random one). We have

$$|Q| = q = 2^t \quad \text{and} \quad |Q'| = \sum_{k \le \gamma t} \binom{t}{k} \approx 2^{H(\gamma)t},$$

where $\gamma = (1+\alpha)^2\gamma_0$.

LEMMA 5.4. *There exists a $[\Delta, R_1\Delta]$ linear q-ary code C_1 such that for sufficiently large Δ and t and any $\beta > 0$,*

$$E_1 \ge \delta_1^q - \sigma' H(\gamma) - 2\beta\sigma.$$

Proof. A nonzero vector falls in a random q-ary linear $[\Delta, R_1\Delta]$ code C_1 with probability equal to $q^{-\Delta}(q^{\Delta R_1} - 1)$. Hence the average number of codewords of weight $i > 0$ in C_1 with at most $\beta\sigma\Delta$ nonzero symbols in $Q \setminus Q'$ equals

$$\mathsf{E}A_{i,\beta} \le q^{-\Delta(1-R_1)} \binom{\Delta}{i} \sum_{j=0}^{\beta\sigma\Delta} \binom{i}{j} |Q'|^{i-j}(q-1)^j.$$

When t is large, by (2.1) we have $1 - R_1 \approx \delta_1^q$, and the exponents of both binomial coefficients are small. Therefore the above inequality can be rewritten as

$$\mathsf{E}A_{i,\beta} \lesssim q^{\Delta(-\delta_1^q + H(\gamma)i + \beta\sigma)}.$$

We now compute the error probability for the right decoder under the condition that the input vector is a random vector of the required pattern. It can be bounded above by the probability that such a vector covers at least half the symbols of a given vector of weight $i > 0$ and of the above pattern, which is not more than

$$q^{\Delta(-H(\gamma)i/2 + \beta\sigma + \varepsilon(\Delta))},$$

where $\varepsilon(\Delta) > 0$ can be made smaller than any given number by an appropriate choice of Δ. The probability that the random error vector covers half a codeword of weight i is not more than

$$A_{i,\beta} q^{\Delta(-H(\gamma)i/2 + \beta\sigma + \varepsilon(\Delta))}. \tag{5.7}$$

As usual, we can choose a code C_1 such that every $A_{i,\beta}$ is not more than $\mathsf{E}A_{i,\beta}$ times a polynomial in n. We obtain therefore that the maximum of (5.7) is obtained when i is as large as possible, namely, $i = 2\sigma'\Delta$.

Switching to exponents we obtain

$$E_1 \ge \delta_1^q - \sigma' H(\gamma) - 2\beta\sigma. \qquad \square$$

Let us complete the proof of Theorem 5.2. As $R \to \mathcal{C}$, the first term in (5.5) tends to zero, while the second remains bounded away from zero. Together with (5.6) this enables us to claim that the lower bound (5.5) is minimized for

$$\sigma \to \frac{\delta_1^q}{H(\gamma_0)} = \frac{1 - R_1}{H(\gamma_0)} = \frac{R_0 - R}{H(\gamma_0)}. \tag{5.8}$$

Substituting this value of σ into (5.5), we obtain the exponent $F(R,p)$ of the theorem.

6. A further improvement of the exponent. Borrowing from turbo-codes. In this section we show how to modify slightly the code construction to improve Theorem 5.2 to the following theorem.

THEOREM 6.1. *For a given rate R, there exists a polynomial-time family of (replicated, generalized) expander codes of length N, defined up to the orderings of the coordinates of the constituent codes, such that, given a random set ω of orderings of the constituent codes,*

$$P_e(C,p,\omega) \le 2^{-NF(R,p)(1-\varepsilon(R,p))},$$

where $\varepsilon(R,p)$ is a function depending only on R and p such that $\varepsilon(R,p) < 1$ when $R < 1 - H(p)$ and $\varepsilon(R,p) \to 0$ when $R \to 1 - H(p)$ and where

$$F(R,p) = \max_{R<R_0<\mathcal{C}} E_0(R_0,p)\frac{1 - R/R_0}{H(\delta_E(R_0))}.$$

The decoding complexity of these codes is the same as in Theorem 4.3.

The idea is based on the following remark. Every bit of an expander code belongs to *two* constituent codes, C_0 and C_1. Turbo-codes, on the other hand, have the property that only the information bits belong to two constituent codes, while the redundancy bits belong to a single one.

To mimic this structure in the expander code context, let us modify the code construction that we have used in the following way: we keep the same expander graph G as before but append some extra edges; namely, we add $at\Delta$ "dangling" edges to every one of the left vertices of A. By "dangling" we mean that we introduce as many vertices (of degree one) as we introduce edges. The right side of the graph is left untouched; so is the constituent code C_1. The left constituent code C_0 is modified only inasmuch as it is now a randomly chosen code of rate R_0 and length $t\Delta(1 + a)$ (instead of $t\Delta$). We now assign coordinate positions to C_0 in such a way that all information bits are assigned edges of the bipartite graph G: in other words, all the "dangling" edges must correspond to redundancy bits. For this to be possible the number of additional edges must be such that $R_0t\Delta(1 + a) \le t\Delta$: to this end we choose $a = R_0^{-1}(1 - R_0 - \varepsilon)$ so that $R_0t\Delta(1 + a) = t\Delta(1 - \varepsilon)$.

The decoding procedure is hardly modified: the first decoding step is again max-likelihood decoding of C_0. The second decoding step is unchanged. The third decoding step and subsequent unevenly numbered decoding steps are slightly modified in the sense that q-ary max-likelihood decoding is again applied, not to the code C_0, however, but to the shortened code C_0^S obtained from C_0 by throwing away the redundancy bits corresponding to the dangling edges. The code C_0^S has q-ary length Δ and the same dimension as C_0, i.e., rate $R_0(1 + a) = 1 - \varepsilon$. Throughout the rest of the decoding procedure the dangling edges are ignored, and their value is recovered only at the very end if the algorithm has converged (and has therefore recovered all the information bits). An examination of the convergence conditions of section 5.3 shows that everything holds in the modified case, with the exception that δ_0^q must now be understood to refer to the q-ary relative minimum distance, not of C_0 but of the shortened code C_0^S. We will be careful to choose ε to be sufficiently big so that we can again claim that the first term in (5.5) dominates the second term. Nevertheless, this does not stop us from having $\varepsilon \to 0$ when $R \to \mathcal{C}$.

As a result, the conditions for convergence are essentially unchanged, but now the overall rate R of the code is higher. Indeed, the new length is now $N = nt\Delta(1 + a)$,

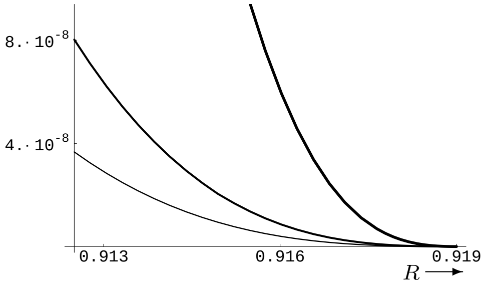


FIG. 6.1. *Error exponents of expander and concatenated codes in the neighborhood of capacity of a BSC with $p = 0.01$, $\mathcal{C} \approx 0.919$. Bottom to top: basic construction of expander codes, Theorem* 5.1*; the exponent E_F of concatenated codes; improved bound of Theorem* 6.1.

and the new redundancy is

$$(1-R)N = n(t\Delta(1-R_1) + t\Delta(1+a)(1-R_0)),$$

which gives, after some rearranging, the expression for the new rate

$$R = R_0R_1 - \frac{\varepsilon}{1-\varepsilon}R_0(1-R_1).$$

Together with (5.5)–(5.8) this implies Theorem 6.1.

We include a sketch of the exponents discussed in the paper in Figure 6.1 for code rate R in the neighborhood of capacity.

7. Concluding comments. For code rates close to capacity the result of Theorem 6.1 improves substantially on our earlier result [2]. Interestingly, for $R \to \mathcal{C}$ we also have $F(R,p) > E_{\mathrm{F}}(R)$; i.e., expander codes have an exponentially smaller error rate than Forney's (quadratic–time-decodable) concatenated codes. A common point that these two families share is the use of two constituent codes; however, in the expander code construction both codes are binary, while in the concatenated scheme the q-ary code C_1 has a strong algebraic structure.

We note that in the study of minimum distance of codes there is a substantial difference between explicit families of codes and randomized ensembles. This difference does not play such a prominent role when we study decoding error exponents. Indeed, in this context the focus is on decoding performance, and the goal is to estimate the error probability of some explicit decoding algorithm. It hardly matters whether this probability is computed for a fixed code and a random error vector or the product of a random code and a random error vector.

Comparing Theorem 6.1 with the error exponent of multilevel concatenations (3.2), we notice that $E_i(R) > F(R,p)$ starting from some finite (not too large) value of i that depends on the channel crossover probability p.

The following research problem suggests itself rather naturally: Is it possible to improve the error exponents of expander codes by using several constituent codes similarly to the improvement (3.2) by Blokh and Zyablov of Forney's exponent $E_{\mathrm{F}}(R)$?

More generally, now that concatenated codes finally have a rival, it should be very interesting to see whether expander codes (or some enlarged family) ultimately have the potential to provide the best constructive exponents. An even more general question is whether there exist polynomial-time decodable code families with error exponent $E(R)$ such that the quotient $E_0(R,p)/E(R)$ stays bounded as $R \to \mathcal{C}$.

Finally, we observe that the study of exponents provides an interesting theoretical framework for trying to discern which properties of codes are relevant for iterative decoding.

Postscriptum of July 8, 2003: During the year-and-a-half that this paper was under review, the following developments took place.

Several ways have been discovered to attain the performance of concatenated codes with expander and expander-like codes, including the full bound $E_{\mathrm{F}}(R)$ [3] and the Zyablov bound on the minimum distance [3], [12].

One of the questions in this section was resolved positively: it is possible to construct and analyze multilevel expander codes [3]. Their error exponents behave in a way similar to those of concatenated codes in (3.2).

Appendix: Proof of Proposition 3.1. We will study properties of an ensemble $\mathcal{A}$ of binary linear codes defined by $(N-K)\times N$ parity-check matrices whose elements are chosen independently with $P(0)=P(1)=1/2$.

Let $K=RN$. We write $f(N)\cong g(N)$ if $\lim_{N\to\infty}\frac{1}{N}\log\frac{f(N)}{g(N)}=0$.

Let $C\in\mathcal{A}$ be a linear code and A_w be the number of vectors of weight w in it. It is easy to see that

$$\mathsf{E}A_w=\binom{N}{w}(2^K-1)2^{-N}.$$

By the Markov inequality there exist codes with

$$A_w\le N\binom{N}{w}2^{K-N}\quad (w=1,2,\dots,N). \tag{7.1}$$

We continue with a technical lemma.

Lemma 7.1. *Let* $S_j=\{x\in\mathcal{H}_2^N : |x|=j\}$, $r\le w\le N/2$,

$$M(w,r):=\sum_{c\in S_w}\bigl|\{y\in S_r : \mathrm{d}(y,c)\le r\}\bigr|.$$

Then for $N\to\infty$, $r=\rho N$, $M(w,r)\lesssim\binom{N}{r}^2$, *with equality*

$$M(w,r)\cong\binom{N}{r}^2$$

only for $\frac{w}{N}\sim 2\rho(1-\rho)$.

Proof.

$$M(w,r)=\binom{N}{w}\sum_{i=w/2}^{r}\binom{w}{i}\binom{N-w}{r-i}.$$

The unconstrained maximum of the summation term is attained when $i=wr/N<$

$w/2$. Hence we write

$$\begin{aligned} M(w,r) &\cong \binom{N}{w}\binom{w}{w/2}\binom{N-w}{r-w/2} \\ &= \binom{N}{r}\binom{N-r}{w/2}\binom{r}{w/2} \qquad \text{(rewriting the multinomial coefficient)} \\ &\lesssim \binom{N}{r}\binom{N-r}{r(1-\rho)}\binom{r}{\rho(N-r)}. \end{aligned}$$

Finding the maximum on w in the last step calls for some explanation. For a fixed vector y of weight r we count the number of weight-w vectors c with $\mathrm{d}(y,c) = r = \rho N$. This number is maximized if c is a typical vector obtained after n independent drawings from the binomial probability distribution given by $P[y+c=1] = \rho, P[y+c=0] = 1-\rho$. Hence the maximizing argument is given by $w/2 = r(1-\rho)$. Substituting it, we obtain

$$\begin{aligned} M(r(1-\rho),r) &\cong \exp[N(H(\rho)+(1-\rho)H(1-\rho)+\rho H(1-\rho))] = \exp(2NH(\rho)) \\ &\cong \binom{N}{r}^2. \qquad \square \end{aligned}$$

Now let us compute the error exponent of maximum likelihood decoding for a code C of rate R with weight distribution as in (7.1). Suppose that $y \in \mathcal{H}^N$ is a vector received from the channel BSC(p) and that the transmitted vector is the all-zero one. Let $d = \delta_{\mathrm{GV}}(R)N$ be the distance of C. Let $\mathcal{E}$ be the event of a decoding error, and let $\mathcal{E}_w \subset \mathcal{E}$ be the event where the decoded codeword is of weight w.

We have

$$\Pr[\mathcal{E}] \le \Sigma_1 + \Sigma_2, \tag{7.2}$$

where

$$\Sigma_1 = \sum_{w=d}^{2d}\sum_{r=d/2}^{d} \Pr\left[\mathcal{E}_w \;\middle|\; |y| = r\right]\Pr[|y| = r], \quad \Sigma_2 = \Pr[|y| \ge d].$$

Now, conditional on the event $|y| = r$, the probability that y is decoded incorrectly to a *given* codeword of weight w equals

$$P_{wr} = \binom{N}{r}^{-1}\sum_{i=w/2}^{r}\binom{w}{i}\binom{N-w}{r-i},$$

and writing

$$\Pr\left[\mathcal{E}_w \;\middle|\; |y| = r\right] \le A_w P_{wr},$$

we get, since $A_w = 2^{-N(1-R)}\binom{N}{w}$ and $\Pr[|y| = r] = \binom{N}{r}p^r(1-p)^{N-r}$,

$$\Sigma_1 \le 2^{-N(1-R)}\sum_{w=d}^{2d}\sum_{r=d/2}^{d} M(w,r)p^r(1-p)^{N-r}.$$

Let $w = \omega N, r = \rho N$. Since we are dealing with exponential terms the sums can asymptotically be replaced with maximums: by Lemma 7.1 the unconstrained maximum on $w = \omega N$ is attained for $\omega = 2\rho(1-\rho)$, and then

$$\begin{aligned}\Sigma_1 &\lesssim 2^{-N(1-R)} \max_{d/2 \le r \le d} \binom{N}{r}^2 p^r (1-p)^{N-r} \\ &= \max_{\delta_{\mathrm{GV}}(R)/2 \le \rho \le \delta_{\mathrm{GV}}(R)} \exp[-N(D(\rho \| p) + (1-R) - H(\rho))].\end{aligned}$$

The unconstrained maximum on ρ on the right-hand side (the minimum of the exponent) is attained for $\rho = \rho_0$ (3.1). We are interested in the case of $R \ge R_{\mathrm{crit}}$, i.e., $\rho_0 \ge \delta_{\mathrm{GV}}(R)$. Then both the present upper bound on Σ_1 and Σ_2 behave (in the $\cong$ sense) as $\exp(-ND(\delta_{\mathrm{GV}}(R)\|p))$. Moreover, the exponent $D(\delta_{\mathrm{GV}}(R)\|p)$ is attained for $\rho = \delta_{\mathrm{GV}}(R)$, and hence for $\omega = \delta_E(R)$.

To prove that $\delta_E(R)$ is indeed the typical relative weight of incorrectly decoded codewords it is now enough to argue that $\Pr[\mathcal{E}] \cong \Sigma_1$, i.e., that the estimate (7.2) and all subsequent upper bounds are in fact asymptotic equalities. This is the so-called sphere-packing bound [11, p. 164], which states that for *any* code with the same rate R we must have $\Pr[\mathcal{E}] \ge \Sigma_2$.

So we have proved that

$$\Pr[\mathcal{E}] \cong \Pr[|y| = d] \cong \Pr\left[\mathcal{E}_w \;\middle|\; |y| = d\right] \Pr[|y| = d]$$

for $w = \delta_E N$ and that $\Pr[\mathcal{E}_w \mid |y| = r] \Pr[|y| = r]$ is an exponentially smaller quantity for all w separated from $\delta_E N$ and for all r. In other words,

$$\Pr\left[\mathcal{E}_w \;\middle|\; \mathcal{E}\right] = \frac{\Pr[\mathcal{E}_w]}{\Pr[\mathcal{E}]} \cong \frac{\Pr[\mathcal{E}_w]}{\Pr[|y| = d]} = \frac{\sum_r \Pr\left[\mathcal{E}_w \;\middle|\; |y| = r\right] \Pr[|y| = r]}{\Pr[|y| = d]}$$

is maximum (and $\cong 1$) for $w = \delta_E N$ and exponentially small for $\omega \neq \delta_E$, which is exactly the statement of Proposition 3.1.

REFERENCES

[1] A. Barg, *Complexity issues in coding theory*, in Handbook of Coding Theory, Vol. 1, V. Pless and W. C. Huffman, eds., Elsevier Science, Amsterdam, 1998, pp. 649–754.

[2] A. Barg and G. Zémor, *Error exponents of expander codes*, IEEE Trans. Inform. Theory, 48 (2002), pp. 1725–1729.

[3] A. Barg and G. Zémor, *Concatenated codes: Serial and parallel*, in Proceedings of the 2003 IEEE International Symposium on Information Theory, Yokohama, Japan, 2003, p. 465.

[4] C. Berrou, A. Glavieux, and P. Thitimajshima, *Near Shannon limit error-correcting coding and decoding*, in Proceedings of the IEEE International Conference on Communication (ICC'93), Geneva, Switzerland, 1993, pp. 1064–1070.

[5] E. L. Blokh and V. V. Zyablov, *Linear Concatenated Codes*, Nauka, Moscow, 1982 (in Russian).

[6] L. Decreusefond and G. Zémor, *On the error-correcting capabilities of cycle codes of graphs*, Combin. Probab. Comput., 6 (1997), pp. 27–38.

[7] I. Dumer, *Concatenated codes and their multilevel generalizations*, in Handbook of Coding Theory, Vol. 2, V. Pless and W. C. Huffman, eds., Elsevier Science, Amsterdam, 1998, pp. 1911–1988.

[8] P. Elias, *Coding for noisy channels*, in IRE Conv. Rec., 1955, pp. 37–46. Reprinted in Key Papers in the Development of Information Theory, D. Slepian, ed., IEEE Press, New York, 1974, pp. 102–111.

[9] G. D. Forney, Jr., *Concantenated Codes*, MIT Press, Cambridge, MA, 1966.

[10] R. G. Gallager, *Low-Density Parity-Check Codes*, MIT Press, Cambridge, MA, 1963.

[11] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley and Sons, New York, 1968.

[12] V. Guruswami and P. Indyk, *Near-optimal linear-time codes for unique decoding and new list-decodable codes over smaller alphabets*, in Proceedings of the 34th ACM Symposium on the Theory of Computing (STOC'02), Montreal, QC, Canada, 2002, pp. 812–821.

[13] A. Lubotzky, R. Phillips, and P. Sarnak, *Ramanujan graphs*, Combinatorica, 8 (1988), pp. 261–277.

[14] G. A. Margulis, *Explicit constructions of graphs without short cycles and low density codes*, Combinatorica, 2 (1982), pp. 71–78.

[15] M. Sipser and D. A. Spielman, *Expander codes*, IEEE Trans. Inform. Theory, 42 (1996), pp. 1710–1722.

[16] M. Tanner, *A recursive approach to low-complexity codes*, IEEE Trans. Inform. Theory, 27 (1981), pp. 533–547.

[17] G. Zémor, *On expander codes*, IEEE Trans. Inform. Theory, 47 (2001), pp. 835–837.

SIAM J. Discrete Math.
Vol. 17, No. 3, pp. 446–452

EQUIVALENCE OF THE 1-RATE MODEL TO THE CLASSICAL MODEL ON STRICTLY NONBLOCKING SWITCHING NETWORKS[*]

W. R. CHEN[†], F. K. HWANG[†], and XUDING ZHU[‡]

Abstract. In the 1-rate(f) network, each link can carry up to f messages for some integer f. The classical model is the special case when $f = 1$. We show that a network is strictly nonblocking under the 1-rate(f) model if and only if it is strictly nonblocking under the classical model.

Key words. switching network, 1-rate network, multirate network, graph coloring, flow, strictly nonblocking

AMS subject classifications. 68M10, 15C15, 90B18

DOI. 10.1137/S0895480102414806

1. Introduction.

A switching network consists of a set of nodes and a set of (directed) links. Typically, an outlink of a node is the inlink of another node, and vice versa. There are two special types of nodes: the *inputs* and the *outputs*. Each input (output) is a node which has no inlink (outlink) and exactly one outlink (inlink).

We view a network as a directed graph $G = (V, E)$, where each vertex is a node and each edge is a link. The inputs and outputs are subsets I, O of V. To emphasize the special roles of the inputs and outputs, we denote a network as $G = (V, E, I, O)$. A network is called *acyclic* if the directed graph G is acyclic; i.e., G contains no directed cycles.

Let $G = (V, E, I, O)$ and f be a positive integer. The 1-rate(f) network, denoted by (G, f), is a network G together with the capacity constraint that each edge can carry up to f messages. If $f = 1$, then the 1-rate network $(G, 1)$ is the *classical model*. In other words, a classical model is a network in which each edge can carry at most one message. In this paper, we consider only 1-rate networks.

A *traffic* of (G, f) is a sequence of input-output pairs (i, j), where $i \in I$ and $j \in O$. There are two types of traffics: requests and cancellations. A *request* is a pair (i, j) such that neither of i, j has appeared in more than $f - 1$ previous uncancelled requests. Namely, the pair requests a connection in the network. A *cancellation* is a previous request whose connection in the network is to be removed. A request (i, j) is *routed* if a directed i-j-path is chosen, without exceeding the capacity of the edges. So a request (i, j) can be routed in the network (which has already routed many previous requests) if and only if there exists a directed i-j-path, each of whose edges has not been used more than $f - 1$ times.

A *state* S of (G, f) is a collection of (not necessarily distinct) directed paths of G joining vertices of I to vertices of O such that each edge e is contained in at most f directed paths. Given a state S, let $S(e)$ denote the number of directed paths

[*] Received by the editors September 19, 2002; accepted for publication (in revised form) July 28, 2003; published electronically January 22, 2004.

http://www.siam.org/journals/sidma/17-3/41480.html

[†] Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan (fhwang@math.nctu.edu.tw). The research of these authors was partially supported by ROC National Science Council grant NSC 91-2115-M-009-010 and the National Chiao Tung University Lee-MTI Center.

[‡] Department of Applied Mathematics, National Sun Yat-sen University, Kaoshiung 80424, Taiwan (zhu@math.nsysu.edu.tw). This author's research was partially supported by ROC National Science Council grant NSC 91-2115-M-110-003.

containing e. Then $0 \leq S(e) \leq f$. A state is *blocking* if there exist a vertex $i \in I$ and $j \in O$ such that both i and j are contained in fewer than f directed paths in S, and every directed i-j-path of G contains an edge e with $S(e) = f$. We say that (G, f) is strictly nonblocking if there is no blocking state.

The classical model is, of course, the dominating model in the study of switching networks. Recently, the multirate network has received increasing attention due to the popular attempt to integrate multimedia service into one network. Since the theory of the classical model is well established, it is profitable to ask how much of it can be extended to the multirate model. The 1-rate model is the simplest multirate model but also has its own application. It is used in the *digital symmetrical matrices* in time-space switching [7, 10]. The principle of providing more links between two nodes, known as *statistical line grouping* in [8], was promoted as a major technique to cut down network blocking. On the other hand, strict nonblockingness is one of the most fundamental properties of a switching network. Therefore, asking whether one model implies the other on this property can serve as a natural start to explore the relation between the classical model and the multirate model. In this paper we prove that if $G = (V, E, I, O)$ is an acyclic network, then the strict nonblockingness of a 1-rate network (G, f) is equivalent to that of the classical model $(G, 1)$.

2. Strictly nonblocking for (G, f) implies the same for $(G, 1)$. We first prove the implication in one direction.

THEOREM 1. *If (G, f) is strictly nonblocking for some positive integer f, then $(G, 1)$ is strictly nonblocking.*

Proof. It suffices to prove that if $(G, 1)$ has a blocking state, then (G, f) has a blocking state. Suppose S is a blocking state of $(G, 1)$. Let S' be the collection of directed paths of G which is obtained by duplicating f times each directed path of S. Then S' is a state of (G, f) and for each edge e of G, $S'(e) = f \times S(e)$. As S is a blocking state of $(G, 1)$, there is an input $i \in I$ and an output $j \in O$ such that none of i, j is contained in any directed path of S, and any directed i-j-path of G contains an edge e with $S(e) = 1$. Then both of i and j are contained in no directed paths of S', and every directed i-j-path of G contains an edge e with $S'(e) = f$. Therefore S' is a blocking state of (G, f). □

In the remainder of this paper, we shall prove the other direction; i.e., if for some integer $f \geq 1$, (G, f) has a blocking state, then $(G, 1)$ has a blocking state. Let S be a blocking state of (G, f). Then there exist $i \in I$ and $j \in O$ such that both i, j are contained in at most $f - 1$ directed path of S, and any directed i-j-path contains an edge e with $S(e) = f$. We need to construct a blocking state S' for $(G, 1)$. One may attempt to partition the directed paths in S into f classes such that

(i) directed paths which share an edge belong to different classes;

(ii) there exists a class C not containing any directed path with end vertex i or j.

If such a partition exists, then it is easy to verify that the class C is a blocking state of $(G, 1)$. However, such a partition may not exist. Consider the following network: Figure 1 shows an example of $(G, 2)$, where G is a simple digraph (a pair of double links indicates a link carrying two paths). The collection of directed paths $S = \{P_1, P_2, P_3, P_4\}$ in Figure 1 is a blocking state for $(G, 2)$, where input i and output j each has generated one path, and hence a new request (i, j) is legitimate. However, it is impossible to partition the paths into two classes in such a way that directed paths sharing an edge belong to different classes, because every two directed paths share an edge. Thus to construct the blocking state S' for $(G, 1)$, we need to use directed paths not contained in the collection S.

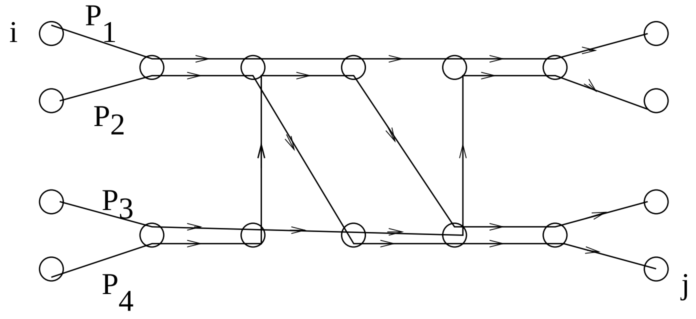


FIG. 1. *An example.*

3. Strictly nonblocking for $(G, 1)$ implies the same for $(G, 2)$.

In this section, we consider the case $f = 2$.

THEOREM 2. *Suppose G is acyclic. If $(G, 1)$ is nonblocking, then $(G, 2)$ is nonblocking.*

Proof. Let S be a blocking state for $(G, 2)$. Thus there exist $i \in I$ and $j \in O$ such that both i, j are contained in at most one directed path of S, and any directed i-j-path contains an edge e with $S(e) = 2$.

We shall construct a blocking state for $(G, 1)$. For each vertex v of G, denote by $E^+(v)$ the outlinks of v and by $E^-(v)$ the inlinks of v. Let $E(v) = E^+(v) \cup E^-(v)$. Let

$$s^+(v) = \sum_{e \in E^+(v)} S(e) = \sum_{P \in S} |P \cap E^+(v)|,$$

$$s^-(v) = \sum_{e \in E^-(v)} S(e) = \sum_{P \in S} |P \cap E^-(v)|,$$

and

$$s(v) = s^+(v) + s^-(v) = \sum_{P \in S} |P \cap E(v)|.$$

Since each directed path $P \in S$ connects a vertex of I to a vertex of O, we conclude that for each vertex $v \notin I \cup O$, $|P \cap E^+(v)| = |P \cap E^-(v)|$. Hence $s^+(v) = s^-(v)$ and $s(v) = 2s^+(v)$. Let $E_1 = \{e \in E : S(e) = 1\}$, and let $E_2 = \{e \in E : S(e) = 2\}$. Then $s(v) = |E_1 \cap E(v)| + 2|E_2 \cap E(v)|$. If $v \notin (I \cup O)$, then $s(v)$ is even, and hence $|E_1 \cap E(v)|$ is even. Let $G_1 = (V, E_1)$ be the subgraph of G induced by the edge set E_1. As each vertex of $V - (I \cup O)$ has even degree in G_1, we can decompose G_1 into an edge-disjoint union of (not necessarily directed) cycles and paths, say

$$E_1 = (P_1 \cup P_2 \cup \cdots \cup P_l) \cup (C_1 \cup C_2 \cup \cdots \cup C_m),$$

where each path P_k connects two vertices of $I \cup O$. We color the edges of each P_k and C_l by two colors, a and b, as described below.

Given an undirected cycle (or a path), there are two choices for the *positive direction* of the cycle (or path). If the cycle is drawn on the plane, then either the clockwise direction or the counterclockwise direction can be chosen as the positive direction. For a path with end vertices i' and j', one can traverse the path from i' to j' or from j' to i'. Once a positive direction is chosen, then those directed edges that agree with the positive direction of the cycle (or path) are called *forward edges*, and those directed edges that oppose the positive direction are called *backward edges.* Arbitrarily choose a positive direction of C_l (or P_k) and color the forward edges of C_l (or P_k) by color a and backward edges by color b, except that if there exist an edge incident to i and an edge incident to j, then they should both be colored a. Observe that if these two edges are contained in a same path, then it is easy to see that they are in the same direction. Therefore whether the two edges are contained in a same path or contained in two distinct paths, by appropriately choosing the positive directions of the paths, they are both forward edges. So the required coloring exists.

Let $E_a \subset E_1$ be the edges of color a and $E_b \subset E_1$ be the edges of color b. Let $B_1 = E_a \cup E_2$ and $B_2 = E_b \cup E_2$. Suppose $v \notin (I \cup O)$. Let $i_a(v)$ (respectively, $o_a(v)$) be the number of inlinks (respectively, outlinks) of v of color a, and let $i_b(v)$ (respectively, $o_b(v)$) be the number of inlinks (respectively, outlinks) of v of color b.

If P_k (or C_l) contains v, then either P_k (or C_l) contains two inlinks or two outlinks of v which are colored by distinct colors or it contains one outlink and one inlink of v which are colored by the same color. Therefore

$$i_a(v) + o_b(v) = o_a(v) + i_b(v).$$

Let $i_2(v) = |E_2 \cap E^-(v)|$ and $o_2(v) = |E_2 \cap E^+(v)|$. Then

$$s^-(v) = i_a(v) + i_b(v) + 2i_2(v)$$

and

$$s^+(v) = o_a(v) + o_b(v) + 2o_2(v).$$

As $s^+(v) = s^-(v)$, we conclude that $i_a(v) + i_2(v) = o_a(v) + o_2(v)$ and $i_b(v) + i_2(v) = o_b(v) + o_2(v)$.

Let H_1 be the directed subgraph of G induced by the edge set $E_a \cup E_2$ and H_2 be the directed subgraph of G induced by the edge set $E_b \cup E_2$. Then for each vertex $v \notin (I \cup O)$, the number of inlinks of v in H_1 is $i_a(v) + i_2(v)$, and the number of outlinks of v in H_1 is $o_a(v) + o_2(v)$. So the number of inlinks of v is equal to the number of outlinks of v. As G is acyclic, H_1 is acyclic. Therefore H_1, and similarly H_2, can be decomposed into directed paths joining vertices of I to vertices of O. For $k = 1, 2$, denote by S_k the collection of directed paths which form a decomposition of H_k. For each edge e of G, $0 \leq S_k(e) \leq 1$ and $S(e) = S_1(e) + S_2(e)$. Moreover, both i and j are not contained in any directed paths of S_2. Any directed i-j-path of G contains an edge e with $S(e) = 2$, and hence $S_2(e) = 1$. Therefore S_2 is a blocking state of $(G, 1)$. □

4. Strictly nonblocking for $(G, 1)$ implies the same for (G, f).

In this section, we prove that the strict nonblocking of the classical model implies the strict nonblocking of the 1-rate(f) model for any $f \geq 1$. Our proof needs a result concerning integer flows of graphs.

Let G be a directed graph. An *integer flow* of G is a mapping $\phi : E \to Z$ which assigns to each edge $e \in E$ an integer $\phi(e)$ such that for each vertex v of G,

$$\sum_{e \in E^+(v)} \phi(e) = \sum_{e \in E^-(v)} \phi(e).$$

An integer flow ϕ is called a *nonnegative* k-flow if for each edge e, $0 \leq \phi(e) \leq k-1$. Lemma 1 is due to Little, Tutte, and Younger [9].

LEMMA 1. *For each nonnegative k-flow f of G, there exist $k-1$ nonnegative 2-flows ϕ_t $(t = 1, 2, \ldots, k-1)$ such that $\phi = \sum_{t=1}^{k-1} \phi_t$.*

LEMMA 2. *Suppose G is acyclic. If S is a state of (G, f), then there are f states $S_1, S_2, \ldots, S_f$ of $(G, 1)$ such that for each edge e of G, $S(e) = \sum_{i=1}^{f} S_i(e)$.*

Proof. Let S be a state of (G, f). Let G' be the directed graph obtained from G by identifying all the inputs and outputs, i.e., identifying all the vertices of $I \cup O$ into a single vertex v^*. We view S as a weight assignment to the edges of G'. It is easy to see that for each vertex v of G',

$$\sum_{e \in E^+(v)} S(e) = \sum_{e \in E^-(v)} S(e),$$

and for each edge of G',

$$0 \leq S(e) \leq f.$$

Therefore S is a nonnegative $(f+1)$-flow of G'. By Lemma 1, G' has f nonnegative 2-flows S_t $(t = 1, 2, \ldots, f)$ such that $S = \sum_{t=1}^{k-1} S_t$. Each nonnegative 2-flow S_t corresponds to a collection of edge disjoint directed cycles of G'. As G is acyclic, each directed cycle C contains the vertex v^*. In other words, each directed cycle C corresponds to a directed path of G joining a vertex of I to a vertex of O. Thus each S_t is indeed a state of $(G, 1)$. □

THEOREM 3. *If $(G, 1)$ is strictly nonblocking, then (G, f) is strictly nonblocking for any $f \geq 1$.*

Proof. Assume (G, f) is not strictly nonblocking and S is a blocking state of (G, f). Then there exist $i \in I$ and $j \in O$ such that both i and j are contained in fewer than f directed paths in S, and every directed i-j-path of G contains an edge e with $S(e) = f$. By Lemma 2, there exist f states, $S_1, S_2, \ldots, S_f$, of $(G, 1)$ such that for every edge e,

$$S(e) = \sum_{k=1}^{f} S_k(e).$$

As both i and j are contained in fewer than f directed paths in S, there exists $1 \leq a, b \leq f$ such that i is not contained in any path of S_a, and j is not contained in any path of S_b. If $a = b$, then S_a is a blocking state of $(S, 1)$. Assume $a \neq b$. Then $S_a \cup S_b$ is a blocking state of $(G, 2)$. By Theorem 2, $(G, 1)$ has a blocking state. □

COROLLARY 1. *Suppose $G = (V, E, O, I)$ is an acyclic network. Then for any positive integers f, f', (G, f) is strictly nonblocking if and only if (G, f') is strictly nonblocking.*

Proof. The strictly nonblocking of (G, f) is equivalent to the strictly nonblocking of $(G, 1)$ for any integer f. Hence strictly nonblocking of (G, f) is equivalent to strictly nonblocking of (G, f'). □

5. Some concluding remarks. Some other implications between the classical model and the multirate model are available from the literature. These involve some other notions of nonblockingness. A network is *wide-sense nonblocking* if every request can be routed, provided all routing follows a given algorithm. A network is *rearrangeably nonblocking* if all requests can be routed if they are given at once (instead of the usual "sequential" model).

Let $C(n_1, r_1, m, n_2, r_2)$ denote the 3-stage Clos network whose nodes are partitioned into three stages (parts):

The first stage consists of r_1 nodes each with n_1 inlinks and m outlinks; the second stage consists of m nodes each with r_1 inlinks and r_2 outlinks; and the third stage consists of r_2 nodes each with m inlinks and n_2 outlinks such that there exists a link from each stage-i node to each stage-$(i+1)$ node but no other links between two nodes.

Clos [4] proved the following lemma.

LEMMA 3. *$C(n_1, r_1, m, n_2, r_2)$ is strictly nonblocking under the classical model if and only if*

$$m \geq \min\{n_1 + n_2 - 1, n_1 r_1, n_2 r_2\}.$$

Hwang and Yeh, as reported in [6], proved a similar result under a model slightly more general than the 1-rate(f) model; suppose each input has capacity f_0, each output has capacity $f_0^{'}$, each link between stage 1 and stage 2 has capacity f_1, and each link between stage 2 and stage 3 has capacity f_2.

LEMMA 4. *$C(n_1, r_1, m, n_2, r_2; f_0, f_0^{'}, f_1, f_2)$ is strictly nonblocking if and only if*

$$m \geq \left\lfloor \frac{\min\{n_1 f_1, n_2 r_2 f_2\} - 1}{f_0} \right\rfloor + \left\lfloor \frac{\min\{n_1 r_1 f_1, n_2 r_2\} - 1}{f_0} \right\rfloor + 1.$$

By setting $f_0 = f_0^{'} = f_1 = f_2 = f$, we obtain the following corollary.

COROLLARY 2. *$C(n_1, r_1, m, n_2, r_2)$ is strictly nonblocking under the 1-rate(f) model if and only if*

$$m \geq \min\{n_1 + n_2 - 1, n_1 r_1, n_2 r_2\}.$$

Note that the conditions in Lemmas 3 and Corollary 2 are the same. Hence we obtain the following theorem.

THEOREM 4. *For $C(n_1, r_1, m, n_2, r_2)$, strictly nonblocking under the classical model implies the same for the 1-rate(f) model, and vice versa.*

Benes [1] proved the following lemma.

LEMMA 5. *$C(n, 2, m, n, 2)$ is wide-sense nonblocking under the classical model if and only if $m \geq \lfloor \frac{3n}{2} \rfloor$.*

On the other hand, Fishburn et al. [5] proved the following lemma.

LEMMA 6. *$C(n, 2, m, n, 2)$ is wide-sense nonblocking under the 1-rate(f) model if and only if $m \geq \lceil \frac{3n}{2} \rceil$.*

By comparing Lemmas 5 and 6, we obtain the following theorem.

THEOREM 5. *For $C(n, 2, m, n, 2)$, wide-sense nonblocking under the classical model does not imply the same for the 1-rate model.*

Finally, Chung and Ross [3] proved the following lemma.

LEMMA 7. *Rearrangeably nonblocking under the classical model implies the same for the 1-rate(f) model.*

For the other direction, only special cases have been proved. Slepian (see [1]) proved the following result (he ignored the terms $n_1 r_1$ and $n_2 r_2$, which reflect the boundary effects).

LEMMA 8. *$C(n_1, r_1, m, n_2, r_2)$ is rearrangeably nonblocking under the classical model if and only if $m \geq \max\{\min\{n_1, n_2 r_2\}, \min\{n_1 r_1, n_2\}\}$.*

On the other hand, Hwang and Yeh, as reported in [6], proved the following lemma.

Lemma 9. *$C(n_1, r_1, m, n_2, r_2; f_0, f_0^{'}, f_1, f_2)$ is rearrangeably nonblocking if and only if*

$$m \geq \max\left\{\frac{\min\{n_1 f_1, n_2 r_2 f_2\}}{f_0}, \frac{\min\{n_1 r_1 f_1, n_2 f_2\}}{f_0^{'}}\right\}.$$

By setting $f_0 = f_0^{'} = f_1 = f_2$, we obtain the following corollary.

Corollary 3. *$C(n_1, r_1, m, n_2, r_2)$ is rearrangeably nonblocking under the 1-rate(f) model if and only if $m \geq \max\{\min\{n_1, n_2 r_2\}, \min\{n_1 r_1, n_2\}\}$.*

By comparing Lemma 8 and Corollary 3, we obtain the following theorem.

Theorem 6. *For the 3-stage Clos network, rearrangeably nonblocking under the 1-rate(f) model implies the same for the classical model.*

Note that all these results deal with the very special 3-stage Clos networks. Chung and Ross, and the authors, are the only exceptions to attack the much harder general networks.

To summarize, we have

	classical		1-rate(f)	remark
strict		$\Longrightarrow$		proved
		$\Longleftarrow$		proved
wide-sense		$\Longrightarrow$		not true
		$\Longleftarrow$		possible
rearrangeable		$\Longrightarrow$		proved
		$\Longleftarrow$		possible

REFERENCES

[1] V. E. Benes, *Mathematical Theory of Connecting Networks and Telephone Traffic*, Academic Press, New York, 1965.

[2] J. A. Bondy and U. S. Murty, *Graph Theory with Applications*, Macmillan, London, 1976.

[3] S.-P. Chung and K. W. Ross, *On nonblocking multirate interconnection networks*, SIAM J. Comput., 20 (1991) pp. 726–736.

[4] C. Clos, *A study of nonblocking switching networks*, Bell System Tech. J., 32 (1953), pp. 406–424.

[5] P. C. Fishburn, F. K. Hwang, D. Z. Du, and B. Gao, *On 1-rate wide-sense nonblocking for 3-stage Clos networks*, Discrete Appl. Math., 78 (1997), pp. 75–87.

[6] F. K. Hwang, *The Mathematical Theory of Nonblocking Switching Network*, World Scientific, Singapore, 1998.

[7] A. Jajszczyk, *On nonblocking switching networks composed of digital symmetrical matrices*, IEEE Trans. Commun., 31 (1983), pp. 2–9.

[8] S.-Y. R. Li, *Algebraic Switching Theory and Broadband Applications*, Academic Press, New York, 2000.

[9] C. H. C. Little, W. T. Tutte, and D. H. Younger, *A theorem on integer flows*, Ars Combin., 26A (1988), pp. 109–112.

[10] P. Odlyzko and S. Das, *Nonblocking rearrangeable networks with distributed control*, in Proceedings of the IEEE International Conference on Communications (ICC'84), Vol. 1, Amsterdam, The Netherlands, 1984, pp. 294–298.

[11] C. Q. Zhang, *Integer Flows and Cycle Double Covers of Graphs*, Marcel Dekker, New York, 1997.

SIAM J. Discrete Math.
Vol. 17, No. 3, pp. 453–477

CONSTRAINT SATISFACTION PROBLEMS ON INTERVALS AND LENGTHS[*]

ANDREI KROKHIN[†], PETER JEAVONS[‡], AND PETER JONSSON[§]

Abstract. We study interval-valued constraint satisfaction problems (CSPs), in which the aim is to find an assignment of intervals to a given set of variables subject to constraints on the relative positions of intervals. Many well-known problems such as Interval Graph Recognition and Interval Satisfiability can be considered as examples of such CSPs. One interesting question concerning such problems is to determine exactly how the complexity of an interval-valued CSP depends on the set of constraints allowed in instances. For the framework known as Allen's interval algebra this question was completely answered earlier by the authors, by giving a complete description of the tractable cases and showing that all remaining cases are NP-complete.

Here we extend the qualitative framework of Allen's algebra with additional constraints on the lengths of intervals. We allow these length constraints to be expressed as Horn disjunctive linear relations, a well-known tractable and sufficiently expressive form of constraints. The class of problems we consider contains, in particular, problems that are very closely related to the previously studied Unit Interval Graph Sandwich problem. We completely characterize sets of qualitative relations for which the CSP augmented with arbitrary length constraints of the above form is tractable. We also show that, again, all the remaining cases are NP-complete.

Key words. Allen's interval algebra, interval satisfiability, computational complexity, tractable cases, dichotomy theorem

AMS subject classifications. 68Q25, 68T20

DOI. 10.1137/S0895480102410201

1. Introduction and summary of results.

A wide range of combinatorial search problems encountered in computer science and artificial intelligence can be naturally expressed as "constraint satisfaction problems" [29], in which the aim is to find an assignment of *values* to a given set of *variables* subject to specified *constraints*. For example, the standard propositional Satisfiability problem [11] may be viewed as a constraint satisfaction problem (CSP) where the variables must be assigned Boolean values, and the constraints are specified by clauses. Further examples include Graph Colorability, Clique, and Bandwidth problems, scheduling problems, and many others (see [2, 19]).

Constraints are usually specified by means of relations. Hence the general CSP can be parameterized according to the relations allowed in an instance. For any set of relations $\mathcal{F}$, the class of CSP instances where the constraint relations are all members of $\mathcal{F}$ is denoted $CSP(\mathcal{F})$. The most well-known examples of such parameterized

[*] Received by the editors June 28, 2002; accepted for publication (in revised form) July 30, 2003; published electronically January 22, 2004. An earlier version of some parts of this paper was published in *Proceedings of the 19th International Symposium on Theoretical Aspects of Computer Science*, Antibes Juan-les-Pins, France, 2002.

http://www.siam.org/journals/sidma/17-3/41020.html

[†] Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK (Andrei.Krokhin@dcs.warwick.ac.uk). The work of this author was partially supported by the UK EPSRC under grant GR/R29598.

[‡] Computing Laboratory, University of Oxford, Oxford OX1 3QD, UK (Peter.Jeavons@comlab.ox.ac.uk). The work of this author was partially supported by the UK EPSRC under grant GR/R29598.

[§] Department of Computer and Information Science, Linköping University, S-581 83 Linköping, Sweden (Peter.Jonsson@ida.liu.se). The work of this author was partially supported by the Swedish Research Council for Engineering Sciences (TFR) under grant 2000-361.

problems are GENERALIZED SATISFIABILITY [34], where the parameter is the set of allowed logical relations, and GRAPH H-COLORING [17], where the parameter is the single graph H.

In studying CSPs over infinite sets of values, arguably the most important type of problem is when the constraints are specified by binary relations and the set of possible values for the variables is the set of intervals on the real line. Such problems arise, for example, in many forms of temporal reasoning [1, 16, 25, 30], where an event is identified with the interval during which it occurs. They also arise in computational biology, where various problems connected with physical mapping of DNA lead to interval-valued constraints [4, 12, 13, 21]. Interval-valued CSPs can be naturally augmented with constraints on the lengths of the intervals, and the complexity of such extended problems will be our main interest in this paper.

Before we describe our new results, we first discuss four closely related families of problems involving intervals which have previously been studied.

The prototypical problem from the first family is the INTERVAL GRAPH RECOGNITION problem [18]. An *interval graph* is an undirected graph such that there is an assignment of intervals to the nodes with two nodes adjacent if and only if the two corresponding intervals intersect. Given an arbitrary graph G, the question of deciding whether G is an interval graph is rarely viewed as a CSP, but in fact it is easily formulated as such a problem in the following way: every pair of adjacent nodes is constrained by the relation $r =$"intersect" over pairs of intervals, and every pair of nonadjacent nodes is constrained by the complementary relation $\bar{r} =$"disjoint." This fundamental INTERVAL GRAPH RECOGNITION problem is tractable, and it also remains tractable if we impose additional constraints on the lengths of the intervals which require all intervals to be of the same length (the UNIT INTERVAL GRAPH RECOGNITION problem [5]). In contrast, it was shown in [32] that if we allow boundaries to be specified for the lengths of intervals, or even exact lengths (which are not necessarily all equal), then the corresponding problems (called BOUNDED INTERVAL GRAPH RECOGNITION and MEASURED INTERVAL GRAPH RECOGNITION, respectively) are NP-complete.

A number of other problems are closely related to the INTERVAL GRAPH RECOGNITION problem, including the CIRCLE GRAPH RECOGNITION problem and the CONTAINMENT GRAPH RECOGNITION problem [10, 15]. These problems can also be formulated as CSPs in a similar way by simply using a different constraint relation.

A typical problem from the second family is the INTERVAL GRAPH SANDWICH problem [13, 16]. Given two graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ such that $E_1 \subseteq E_2$, the question is whether there is an interval graph $G = (V, E)$ with $E_1 \subseteq E \subseteq E_2$. Clearly, this is a generalization of the corresponding recognition problem (the case when $E_1 = E_2$). The INTERVAL GRAPH SANDWICH problem can be represented as a CSP as follows: to any $e \in E_1$ assign the constraint $r =$"intersect," to any $e \notin E_2$ assign the constraint $\bar{r} =$"disjoint," and leave all pairs of variables corresponding to edges from $E_2 \setminus E_1$ unrelated. This problem was shown to be NP-complete along with the UNIT INTERVAL GRAPH SANDWICH problem, where all intervals are required to be of the same length [13].

GRAPH SANDWICH problems for a variety of other graph properties have also been considered [14]. For example, the CIRCLE GRAPH SANDWICH problem is obtained from the INTERVAL GRAPH SANDWICH problem by changing "interval graph G" to "circle graph G." This problem was shown to be NP-complete in [14]; it can be formulated as a CSP in the same way as above using the constraint relation

$r =$“overlap.”

The third family of problems we mention is the so-called INTERVAL SATISFIABILITY problems [16, 33, 35]. In these problems every pair of interval variables is again constrained in some way, but the constraints this time are chosen from a given set $\mathcal{F}$ of relations. In [16, 33, 35] only a small number of possibilities for $\mathcal{F}$ are considered. It is shown there that for some choices of $\mathcal{F}$ the resulting problem is tractable, while for others it is NP-complete. The complexity of INTERVAL SATISFIABILITY with all intervals of the same length is also studied in [33].

The fourth type of problem we mention is the satisfiability problem for *Allen's interval algebra* [1], denoted $\mathcal{A}$-SAT. Allen's algebra contains 13 basic relations (corresponding to the 13 ways two given intervals can be related from a “qualitative” point of view). The set $\mathcal{A}$ contains not just these basic relations but all $2^{13} = 8192$ possible unions of them. The problems $\mathcal{A}$-SAT($\mathcal{F}$) are similar to problems of the third type above, except that not every pair of pair of variables has to be constrained. They can also be represented as INTERVAL SATISFIABILITY with $\mathcal{F}$ being an arbitrary subset of $\mathcal{A}$ containing the total relation. The complexity of problems of the form $\mathcal{A}$-SAT($\mathcal{F}$) has been intensively studied in the artificial intelligence community (see, e.g., [7, 8, 30]), and a complete classification of the complexity of such problems was obtained in [25]. In that paper it is shown that there are exactly 18 maximal tractable fragments of $\mathcal{A}$, and for any subset $\mathcal{F}$ not entirely contained in one of those the problem $\mathcal{A}$-SAT($\mathcal{F}$) is NP-complete.

Many variants of $\mathcal{A}$-SAT($\mathcal{F}$) where additional constraints are allowed have been considered in the literature; cf. [3, 22, 28]. For instance, certain scheduling problems can conveniently be expressed as $\mathcal{A}$-SAT($\mathcal{F}$) with additional constraints on the lengths of the intervals. Moreover, in [2] it was suggested that many important forms of constraints on lengths can be expressed in the form of Horn disjunctive linear relations. This class of relations is known to be tractable [20] and at the same time allows us to express all elementary constraints, such as fixing the length, bounding the length of an interval by a given number, or comparing the lengths of two intervals. It was proved in [2] that only three out of the 18 maximal tractable fragments for $\mathcal{A}$-SAT($\mathcal{F}$) preserve tractability when extended with Horn disjunctive linear constraints on lengths; the other 15 become NP-complete. In this paper we study how we need to further restrict those 15 fragments to obtain tractable cases. The main result is a complete classification of complexity for $\mathcal{A}$-SAT($\mathcal{F}$) with additional constraints on lengths. We show that such problems are either tractable or strongly NP-complete. Moreover, we give a complete description of the tractable cases, which allows one to easily determine whether a given set $\mathcal{F}$ falls into one of the tractable cases.

As well as giving a complete classification, our result also establishes a new dichotomy theorem for complexity. Dichotomy theorems are results concerning a class of related problems (with some parameter) which assert that, for some values of the parameter, the problems in the class are tractable while for all other values they are NP-complete. Such theorems are of interest because it is well known [26] that if P≠NP, then, within NP, there are infinitely many pairwise inequivalent problems of intermediate complexity. Dichotomy results rule out such a possibility within certain classes of problems.

Dichotomy theorems have previously been established for the GENERALIZED SATISFIABILITY [34] and GRAPH H-COLORING [17] problems mentioned above as well as the DIRECTED SUBGRAPH HOMEOMORPHISM problem [9].

CSPs have been a fruitful source of dichotomy results (see, e.g., [6, 23]). For

Table 1

The 13 basic relations of Allen's interval algebra. (The endpoint relations $x^- < x^+$ and $y^- < y^+$ that are valid for all relations have been omitted.)

Basic relation		Example	Endpoints
x precedes y	p	xxx	$x^+ < y^-$
y preceded by x	p^{-1}	yyy	
x meets y	m	xxxx	$x^+ = y^-$
y met by x	m^{-1}	yyyy	
x overlaps y	o	xxxx	$x^- < y^- < x^+$,
y overlapped by x	o^{-1}	yyyy	$x^+ < y^+$
x during y	d	xxx	$x^- > y^-$,
y includes x	d^{-1}	yyyyyyy	$x^+ < y^+$
x starts y	s	xxx	$x^- = y^-$,
y started by x	s^{-1}	yyyyyyy	$x^+ < y^+$
x finishes y	f	xxx	$x^+ = y^+$,
y finished by x	f^{-1}	yyyyyyy	$x^- > y^-$
x equals y	$\equiv$	xxxx	$x^- = y^-$,
		yyyy	$x^+ = y^+$

CSPs, the relevant parameter is usually the set of relations, $\mathcal{F}$, specifying the allowed constraints. This parameter usually runs over an infinite set of values. In the case of Allen's algebra, even though the number of different values for $\mathcal{F}$ is finite, it is astronomical ($2^{8192} \approx 10^{2466}$), which excludes the possibility of computer-aided exhaustive case analysis.

The usual tool for proving dichotomy theorems is reducibility via expressibility. This is done by showing that one set of relations expresses another so that one problem can be reduced to the other. This is the method used in [6, 25, 34], and a similar method is used here. After identifying certain tractable fragments, we find some NP-complete fragments and then show how any subset not entirely contained in one of the tractable sets can express some already known NP-complete fragment.

2. Preliminaries and background. Allen's interval algebra [1], denoted $\mathcal{A}$, is a formalism for expressing *qualitative* binary relations between intervals on the real line. By "qualitative" we mean "invariant under all continuous injective monotone transformations of the real line." An interval x is represented as a pair $[x^-, x^+]$ of real numbers with $x^- < x^+$, denoting the left and right endpoints of the interval, respectively. The qualitative relations between intervals are the $2^{13} = 8192$ possible unions of the 13 *basic interval relations*, which are shown in Table 1. It is easy to see that the basic relations are jointly exhaustive and pairwise disjoint in the sense that any two given intervals are related by exactly one basic relation. For the sake of brevity, relations between intervals will be written as collections of basic relations, omitting the sign of union. So, for instance, we write (pmf^{-1}) instead of $\mathsf{p} \cup \mathsf{m} \cup \mathsf{f}^{-1}$.

The problem of *satisfiability* ($\mathcal{A}$-SAT) in Allen's algebra is defined as follows.

Definition 2.1. *Let $\mathcal{F} \subseteq \mathcal{A}$ be a set of interval relations. An instance I of $\mathcal{A}$-SAT$(\mathcal{F})$ over a set, V, of variables is a set of constraints of the form xry, where $x, y \in V$ and $r \in \mathcal{F}$. The question is whether I is* satisfiable, *i.e., whether there exists a function, f, from V to the set of all intervals such that $f(x)\, r\, f(y)$ holds for every constraint xry in I. Any such function f is called a* model *of I.*

Example 2.1. The instance $\{x(\mathsf{m})y, y(\mathsf{m})z, x(\mathsf{m})z\}$ is not satisfiable because the first two constraints imply that interval x must precede interval z, which contradicts the third constraint.

Example 2.2. The instance $I = \{x(\mathsf{mo})y, y(\mathsf{df}^{-1})z, z(\equiv \mathsf{pmod}^{-1}\mathsf{ss}^{-1}\mathsf{f}^{-1})x\}$ is

Table 2
Composition table for the basic relations in Allen's algebra.

$\circ$	$\equiv$	p	p^{-1}	m	m^{-1}	o	o^{-1}	d	d^{-1}	s	s^{-1}	f	f^{-1}
$\equiv$	$\equiv$	p	p^{-1}	m	m^{-1}	o	o^{-1}	d	d^{-1}	s	s^{-1}	f	f^{-1}
p	p	p	$\top$	p	ρ	p	ρ	ρ	p	p	p	ρ	p
p^{-1}	p^{-1}	$\top$	p^{-1}	λ^{-1}	p^{-1}	λ^{-1}	p^{-1}	λ^{-1}	p^{-1}	λ^{-1}	p^{-1}	p^{-1}	p^{-1}
m	m	p	ρ^{-1}	p	θ	p	β	β	p	m	m	β	p
m^{-1}	m^{-1}	λ	p^{-1}	σ	p^{-1}	γ^{-1}	p^{-1}	γ^{-1}	p^{-1}	γ^{-1}	p^{-1}	m^{-1}	m^{-1}
o	o	p	ρ^{-1}	p	β^{-1}	α	ν	β	λ	o	γ	β	α
o^{-1}	o^{-1}	λ	p^{-1}	γ	p^{-1}	ν	α^{-1}	γ^{-1}	ρ^{-1}	γ^{-1}	α^{-1}	o^{-1}	β^{-1}
d	d	p	p^{-1}	p	p^{-1}	ρ	λ^{-1}	d	$\top$	d	λ^{-1}	d	ρ
d^{-1}	d^{-1}	λ	ρ^{-1}	γ	β^{-1}	γ	β^{-1}	ν	d^{-1}	γ	d^{-1}	β^{-1}	d^{-1}
s	s	p	p^{-1}	p	m^{-1}	α	γ^{-1}	d	λ	s	σ	d	α
s^{-1}	s^{-1}	λ	p^{-1}	γ	m^{-1}	γ	o^{-1}	γ^{-1}	d^{-1}	σ	s^{-1}	o^{-1}	d^{-1}
f	f	p	p^{-1}	m	p^{-1}	β	α^{-1}	d	ρ^{-1}	d	α^{-1}	f	θ
f^{-1}	f^{-1}	p	ρ^{-1}	m	β^{-1}	o	β^{-1}	β	d^{-1}	o	d^{-1}	θ	f^{-1}

$\alpha = (\mathsf{pmo})$ $\beta = (\mathsf{ods})$ $\gamma = (\mathsf{od}^{-1}\mathsf{f}^{-1})$ $\sigma = (\equiv \mathsf{ss}^{-1})$ $\theta = (\equiv \mathsf{ff}^{-1})$
$\rho = (\mathsf{pmods})$ $\lambda = (\mathsf{pmod}^{-1}\mathsf{f}^{-1})$ $\nu = (\equiv \mathsf{oo}^{-1}\mathsf{dd}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})$
$\top = (\equiv \mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}\mathsf{dd}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})$

satisfiable. The function f given by $f(x) = [0,2]$, $f(y) = [1,3]$, and $f(z) = [0,4]$ is a model of I.

An instance of $\mathcal{A}$-SAT($\mathcal{F}$) can also be represented, in an obvious way, as a labelled digraph, where the nodes are the variables from V, and the labelled arcs correspond to the constraints. This way of representing instances is sometimes more transparent.

Allen's interval algebra $\mathcal{A}$ consists of the 8192 possible relations between intervals together with three standard operations on binary relations: *converse* $\cdot^{-1}$*, intersection* $\cap$*, and composition* $\circ$. It is easy to see that the converse of $r = (b_1 \cdots b_n)$ is equal to $(b_1^{-1} \cdots b_n^{-1})$. Using the definition of composition, it can be shown that

$$(b_1 \cdots b_n) \circ (b'_1 \cdots b'_m) = \bigcup \{b_i \circ b'_j \mid 1 \leq i \leq n,\ 1 \leq j \leq m\}.$$

Hence the composition of two relations $r_1, r_2 \in \mathcal{A}$ is determined by the compositions of the basic relations they contain. The compositions of all possible pairs of basic relations are given in Table 2.

Subsets of $\mathcal{A}$ that are closed under the operations of converse, intersection, and composition are said to be *subalgebras*. For a given subset $\mathcal{F}$ of $\mathcal{A}$, the smallest subalgebra containing $\mathcal{F}$ is called the subalgebra *generated* by $\mathcal{F}$ and is denoted by $\langle \mathcal{F} \rangle$. It is easy to see that $\langle \mathcal{F} \rangle$ is obtained from $\mathcal{F}$ by adding all relations that can be obtained from the relations in $\mathcal{F}$ by using the three operations of the algebra $\mathcal{A}$.

It is known [30] and easy to prove that, for every $\mathcal{F} \subseteq \mathcal{A}$, the problem $\mathcal{A}$-SAT($\langle \mathcal{F} \rangle$) is polynomially equivalent to $\mathcal{A}$-SAT($\mathcal{F}$). Therefore, to classify the complexity of $\mathcal{A}$-SAT($\mathcal{F}$) it is sufficient to consider *subalgebras* of $\mathcal{A}$. Throughout the paper, $\mathcal{S}$ denotes a subalgebra of $\mathcal{A}$.

In the following we shall use the symbol ±, which should be interpreted as follows. A condition involving ± means the conjunction of two conditions: one corresponding

Table 3
The 18 *maximal tractable subalgebras of Allen's algebra.*

$$\mathcal{S}_{\mathsf{p}} = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{p})^{\pm 1} \subseteq r\}$$
$$\mathcal{S}_{\mathsf{d}} = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{d}^{-1})^{\pm 1} \subseteq r\}$$
$$\mathcal{S}_{\mathsf{o}} = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{o})^{\pm 1} \subseteq r\}$$
$$\mathcal{A}_1 = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{s}^{-1})^{\pm 1} \subseteq r\}$$
$$\mathcal{A}_2 = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{s})^{\pm 1} \subseteq r\}$$
$$\mathcal{A}_3 = \{r \mid r \cap (\mathsf{pmodf})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{s})^{\pm 1} \subseteq r\}$$
$$\mathcal{A}_4 = \{r \mid r \cap (\mathsf{pmodf}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{s})^{\pm 1} \subseteq r\}$$

$$\mathcal{E}_{\mathsf{p}} = \{r \mid r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{p})^{\pm 1} \subseteq r\}$$
$$\mathcal{E}_{\mathsf{d}} = \{r \mid r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{d})^{\pm 1} \subseteq r\}$$
$$\mathcal{E}_{\mathsf{o}} = \{r \mid r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{o})^{\pm 1} \subseteq r\}$$
$$\mathcal{B}_1 = \{r \mid r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{f}^{-1})^{\pm 1} \subseteq r\}$$
$$\mathcal{B}_2 = \{r \mid r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{f})^{\pm 1} \subseteq r\}$$
$$\mathcal{B}_3 = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{s}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{f}^{-1})^{\pm 1} \subseteq r\}$$
$$\mathcal{B}_4 = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{s})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{f}^{-1})^{\pm 1} \subseteq r\}$$

$$\mathcal{E}^* = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{s})^{\pm 1} \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{ff}^{-1}) \neq \emptyset \Rightarrow (\equiv) \subseteq r \end{array} \right\}$$

$$\mathcal{S}^* = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \cap (\mathsf{pmod}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{f}^{-1})^{\pm 1} \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{ss}^{-1}) \neq \emptyset \Rightarrow (\equiv) \subseteq r \end{array} \right\}$$

$$\mathcal{H} = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \cap (\mathsf{os})^{\pm 1} \neq \emptyset \ \&\ r \cap (\mathsf{o}^{-1}\mathsf{f})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{d})^{\pm 1} \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{ds})^{\pm 1} \neq \emptyset \ \&\ r \cap (\mathsf{d}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{o})^{\pm 1} \subseteq r, \text{ and} \\ 3)\ r \cap (\mathsf{pm})^{\pm 1} \neq \emptyset \ \&\ r \not\subseteq (\mathsf{pm})^{\pm 1} \Rightarrow (\mathsf{o})^{\pm 1} \subseteq r \end{array} \right\}$$

$$\mathcal{A}_{\equiv} = \{r \mid r \neq \emptyset \Rightarrow (\equiv) \subseteq r\}$$

to + and one corresponding to −. For example, the condition

$$r \cap (\mathsf{dsf})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{d})^{\pm 1} \subseteq r$$

means that both of the following conditions hold:

$$r \cap (\mathsf{dsf}) \neq \emptyset \Rightarrow (\mathsf{d}) \subseteq r,$$
$$r \cap (\mathsf{d}^{-1}\mathsf{s}^{-1}\mathsf{f}^{-1}) \neq \emptyset \Rightarrow (\mathsf{d}^{-1}) \subseteq r.$$

The main advantage of using the ± symbol is conciseness: in any subalgebra of $\mathcal{A}$, the "+" and the "−" conditions are satisfied (or not satisfied) simultaneously, and therefore only one of them needs to be verified.

A complete classification of the complexity of problems of the form $\mathcal{A}$-SAT($\mathcal{F}$) was obtained in [25].

Theorem 2.2 (see [25]). *For any subset $\mathcal{F}$ of $\mathcal{A}$, either $\mathcal{A}$-SAT($\mathcal{F}$) is NP-complete or $\mathcal{F}$ is included in $\mathcal{S}$, where $\mathcal{S}$ is one of the* 18 *subalgebras listed in Table* 3*, for which $\mathcal{A}$-SAT($\mathcal{S}$) is tractable.*

In this paper we present a complete complexity classification for a more general problem, namely, for $\mathcal{A}$-SAT($\mathcal{F}$) extended with constraints on the lengths of intervals. Now we define the exact form of constraints on lengths we shall allow.

Definition 2.3. *Let V be a set of real-valued variables and α, β linear polynomials over V with rational coefficients. A* linear relation *over V is an expression of the form $\alpha R \beta$, where $R \in \{<, \leq, =, \neq, \geq, >\}$.*

A disjunctive linear relation *(DLR) over V is a disjunction of a nonempty finite set of linear relations. A DLR is said to be* Horn *if and only if at most one of its disjuncts is not of the form $\alpha \neq \beta$.*

Example 2.3. The expression

$$x + 2y \leq 3z + 42.3$$

is a linear relation,

$$(x + 2y \leq 3z + 42.3) \vee (x + z < 4y - 8) \vee \left(x > \frac{3}{12}\right)$$

is a disjunctive linear relation, and

$$(x + 2y \leq 3z + 42.3) \vee (x + z \neq 4y - 8) \vee \left(x \neq \frac{3}{12}\right)$$

is a Horn disjunctive linear relation.

Definition 2.4. *The problem of* satisfiability *for finite sets, D, of DLRs, denoted* DLRsat*, is that of checking whether there exists an assignment f from variables in V to real numbers such that all DLRs in D are satisfied. Such an f is said to be a* model *of D. The satisfiability problem for finite sets of Horn DLRs is denoted* hornDLRsat*.*

Theorem 2.5 (see [20, 24]). *The problem* DLRsat *is NP-complete, but the problem* hornDLRsat *is solvable in polynomial time.*

We are interested in how the complexity of a problem depends on the value of the parameter $\mathcal{F}$ which, in our case, is a set of qualitative relations. Therefore we shall allow only those constraints on lengths which can be expressed by Horn DLRs and thus are tractable. This class of constraints subsumes all forms of constraints on lengths which have been considered in [32, 33].

We can now define the general interval satisfiability problem with constraints on lengths.

Definition 2.6. *An instance of the problem of* interval satisfiability with constraints on lengths *for a set $\mathcal{F} \subseteq \mathcal{A}$, denoted $\mathcal{A}^l$-sat$(\mathcal{F})$, is a pair $Q = (I, D)$, where*

(i) *I is an instance of $\mathcal{A}$-sat$(\mathcal{F})$ over a set V of variables and*

(ii) *D is an instance of* hornDLRsat *over the set of variables $\{l(v) \mid v \in V\}$.*

The question is whether Q is satisfiable*, i.e., whether there exists a model f of I such that the DLRs in D are satisfied with $l(v)$ equal to the length of $f(v)$ for all $v \in V$.*

Example 2.4. Consider the instance $Q = (I, D)$, where $I = \{x(\mathsf{mo})y, y(\mathsf{df}^{-1})z, z(\equiv \mathsf{pmod}^{-1}\mathsf{ss}^{-1}\mathsf{f}^{-1})x\}$, as in Example 2.2, and $D = \{l(x) > l(y) + l(z)\}$. This instance is not satisfiable: any set of intervals satisfying the constraints in I must have $z^- \leq x^- < x^+ < y^+$ and $y \cap z$ nonempty and thus cannot satisfy the length constraint in D.

Proposition 2.7. *$\mathcal{A}^l$-sat$(\mathcal{F}) \in NP$ for every $\mathcal{F} \subseteq \mathcal{A}$.*

Proof. Every instance of $\mathcal{A}^l$-sat*$(\mathcal{F})$* over a set of variables V can be translated in a straightforward way into an instance of DLRsat over the set of variables $\{v^-, v^+ \mid v \in V\}$. Now the result follows from Theorem 2.5. □

Example 2.5. The instance $Q = (I, D)$ defined in Example 2.4 corresponds to the instance D' of DLRSAT containing the following constraints:

$$\begin{aligned}
&(x^- < x^+),\\
&(y^- < y^+),\\
&(z^- < z^+),\\
&\left.\begin{aligned}&(x^+ = y^-) \vee (x^- < y^-),\\ &(x^+ = y^-) \vee (y^- < x^+),\\ &(x^+ = y^-) \vee (x^+ < y^+),\end{aligned}\right\}\text{corresponding to } x(\mathsf{mo})y\\
&\left.\begin{aligned}&(y^- > z^-) \vee (y^+ = z^+),\\ &(y^- > z^-) \vee (y^- < z^-),\\ &(y^+ \le z^+),\\ &(y^+ < z^+) \vee (y^- < z^-),\end{aligned}\right\}\text{corresponding to } y(\mathsf{df}^{-1})z\\
&z^- \le x^-, \left.\vphantom{z^-}\right\}\text{corresponding to } z(\equiv \mathsf{pmod}^{-1}\mathsf{ss}^{-1}\mathsf{f}^{-1})x\\
&(x^+ - x^-) > (y^+ - y^-) + (z^+ - z^-).
\end{aligned}$$

The complexity of $\mathcal{A}^l$-SAT*($\mathcal{S}$)* has already been determined for each subalgebra $\mathcal{S}$ identified in Theorem 2.2.

PROPOSITION 2.8 (see [2]). *The problem* $\mathcal{A}^l$-SAT$(\mathcal{S})$ *is tractable for* $\mathcal{S} \in \{\mathcal{S}_\mathsf{p}, \mathcal{E}_\mathsf{p}, \mathcal{H}\}$ *and is NP-complete for the other* 15 *subalgebras listed in Table* 3.

In the next section, we determine the complexity of $\mathcal{A}^l$-SAT*($\mathcal{F}$)* for every possible subset $\mathcal{F} \subseteq \mathcal{A}$.

3. Main result.

THEOREM 3.1. *For any subset* $\mathcal{F}$ *of* $\mathcal{A}$*, either* $\mathcal{A}^l$-SAT*($\mathcal{F}$) is strongly NP-complete or* $\mathcal{F}$ *is included in* $\mathcal{S}$*, where* $\mathcal{S}$ *is one of the* 10 *subalgebras listed in Table* 4*, for which* $\mathcal{A}^l$-SAT*($\mathcal{S}$) is tractable.*

In section 3.1, we discuss polynomial-time algorithms for the 10 subalgebras listed in Table 4, and in section 3.2 we give the NP-completeness results we need. (Strong NP-completeness of the NP-complete cases follows from the fact that the biggest number used in these NP-completeness proofs is 5.) Finally, in section 3.3, we give the classification proof.

The following notation is used throughout the proofs: if f is a model of an instance over a set V of variables and $v \in V$, then we denote the left and right endpoints of $f(v)$ by $f(v^-)$ and $f(v^+)$, respectively.

We shall say that a relation is *nontrivial* if it is not equal to the empty relation or the relation $(\equiv)$. Given a relation $r \in \mathcal{A}$, we write r^* to denote the relation $r \cap r^{-1}$. Evidently, every subalgebra of $\mathcal{A}$ is closed under the operation $\cdot^*$ (of taking the symmetric part of a relation).

Now we introduce the notion of *derivation with lengths* which will be used frequently in the proofs below. This notion is an extension of the notion of derivation in Allen's algebra used in [25].

Suppose $\mathcal{F} \subseteq \mathcal{A}$ and $Q = (I, D)$ is an instance of $\mathcal{A}^l$-SAT$(\mathcal{F})$. Let variables x, y be involved in I. Suppose a relation $r \in \mathcal{A}$ satisfies the following condition: Q is satisfiable if and only if xry. Then we say that r is *derived (with lengths)* from $\mathcal{F}$. It can easily be checked that the problems $\mathcal{A}^l$-SAT$(\mathcal{F})$ and $\mathcal{A}^l$-SAT$(\mathcal{F} \cup \{r\})$ are polynomially equivalent because, in any instance of the second problem, any constraint

Table 4
The 10 *tractable cases of* $\mathcal{A}^l$-SAT.

$$\mathcal{S}_{\mathsf{p}} = \{r \mid r \cap (\mathsf{pmod}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{p})^{\pm 1} \subseteq r\}$$

$$\mathcal{E}_{\mathsf{p}} = \{r \mid r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{p})^{\pm 1} \subseteq r\}$$

$$\mathcal{H} = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \cap (\mathsf{os})^{\pm 1} \neq \emptyset \ \&\ r \cap (\mathsf{o}^{-1}\mathsf{f})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{d})^{\pm 1} \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{ds})^{\pm 1} \neq \emptyset \ \&\ r \cap (\mathsf{d}^{-1}\mathsf{f}^{-1})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{o})^{\pm 1} \subseteq r, \text{ and} \\ 3)\ r \cap (\mathsf{pm})^{\pm 1} \neq \emptyset \ \&\ r \not\subseteq (\mathsf{pm})^{\pm 1} \Rightarrow (\mathsf{o})^{\pm 1} \subseteq r \end{array} \right\}$$

$$\mathcal{C}_{\mathsf{o}} = \{r \mid r \neq \emptyset \Rightarrow (\mathsf{oo}^{-1}) \subseteq r\}$$

$$\mathcal{C}_{\mathsf{m}} = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \neq \emptyset \Rightarrow (\mathsf{mm}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1}) \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{pp}^{-1}\mathsf{oo}^{-1}) \neq \emptyset \Rightarrow (\equiv) \subseteq r \end{array} \right\}$$

$$\mathcal{D}_{\mathsf{s}} = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \cap (\mathsf{dsf})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{s})^{\pm 1} \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}) \neq \emptyset \Rightarrow (\equiv \mathsf{ss}^{-1}) \subseteq r \end{array} \right\}$$

$$\mathcal{D}_{\mathsf{f}} = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \cap (\mathsf{dsf})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{f})^{\pm 1} \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}) \neq \emptyset \Rightarrow (\equiv \mathsf{ff}^{-1}) \subseteq r \end{array} \right\}$$

$$\mathcal{D}_{\mathsf{d}} = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \cap (\mathsf{dsf})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{d})^{\pm 1} \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}) \neq \emptyset \Rightarrow (\equiv \mathsf{dd}^{-1}) \subseteq r \end{array} \right\}$$

$$\mathcal{D}'_{\mathsf{d}} = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \cap (\mathsf{dsf})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{d})^{\pm 1} \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{pmo})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{odd}^{-1})^{\pm 1} \subseteq r \end{array} \right\}$$

$$\mathcal{D}''_{\mathsf{d}} = \left\{ r \;\middle|\; \begin{array}{l} 1)\ r \cap (\mathsf{dsf})^{\pm 1} \neq \emptyset \Rightarrow (\mathsf{d})^{\pm 1} \subseteq r, \text{ and} \\ 2)\ r \cap (\mathsf{pp}^{-1}\mathsf{oo}^{-1}) \neq \emptyset \Rightarrow (\mathsf{oo}^{-1}\mathsf{dd}^{-1}) \subseteq r, \text{ and} \\ 3)\ r \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}) \neq \emptyset \Rightarrow (\equiv \mathsf{dd}^{-1}) \subseteq r \end{array} \right\}$$

involving r can be replaced by the set of constraints in Q (introducing fresh variables as needed), and this can be done in polynomial time. It follows that it is sufficient to classify the complexity of problems $\mathcal{A}^l$-SAT$(\mathcal{S})$, where $\mathcal{S}$ is a subalgebra of $\mathcal{A}$ closed under derivations with lengths.

Note that if we prove that the 10 sets shown in Table 4 are the only maximal sets $\mathcal{F}$ for which $\mathcal{A}^l$-SAT$(\mathcal{F})$ is tractable, then it will follow that they are all subalgebras closed under derivation with lengths; that is, we do not have to give a separate proof of this fact.

We will also use the following principle of *duality* to reduce the number of cases to be considered in the forthcoming proofs. We make use of a function reverse which is defined on the basic relations of $\mathcal{A}$ by the following table:

b	$\equiv$	p	p^{-1}	m	m^{-1}	o	o^{-1}	d	d^{-1}	s	s^{-1}	f	f^{-1}
reverse(b)	$\equiv$	p^{-1}	p	m^{-1}	m	o^{-1}	o	d	d^{-1}	f	f^{-1}	s	s^{-1}

It is also defined for all other elements of $\mathcal{A}$ by setting $\mathsf{reverse}(r) = \bigcup_{b \subseteq r} \mathsf{reverse}(b)$.

Let $Q = (I, D)$ be any instance of $\mathcal{A}^l$-SAT with set of variables V, and let $Q' =$

(I', D) be the instance obtained from Q by replacing every r in I with $\mathsf{reverse}(r)$. It is easy to check that Q has a model f if and only if Q' has a model f' given by

$$f'(v) = [-f(v^+), -f(v^-)] \text{ for all } v \in V.$$

In other words, f' is obtained from f by redirecting the real line and leaving all intervals (as geometric objects) in their places. This observation leads to the following lemma.

LEMMA 3.2. *Let $\mathcal{F} = \{r_1, \dots, r_n\} \subseteq \mathcal{A}$ and $\mathcal{F}' = \{r'_1, \dots, r'_n\} \subseteq \mathcal{A}$ be such that, for all $1 \leq k \leq n$, $r'_k = \mathsf{reverse}(r_k)$. Then $\mathcal{A}^l$-SAT$(\mathcal{F})$ is tractable (NP-complete) if and only if $\mathcal{A}^l$-SAT$(\mathcal{F}')$ is tractable (NP-complete).*

3.1. Tractability results.

PROPOSITION 3.3. *The problem $\mathcal{A}^l$-SAT$(\mathcal{S})$ is tractable whenever $\mathcal{S}$ is one of $\mathcal{S}_{\mathsf{p}}$, $\mathcal{E}_{\mathsf{p}}$, $\mathcal{H}$, $\mathcal{C}_{\mathsf{o}}$, $\mathcal{C}_{\mathsf{m}}$, $\mathcal{D}_{\mathsf{s}}$, $\mathcal{D}_{\mathsf{f}}$, $\mathcal{D}_{\mathsf{d}}$, $\mathcal{D}'_{\mathsf{d}}$, or $\mathcal{D}''_{\mathsf{d}}$.*

Polynomial-time algorithms solving $\mathcal{A}^l$-SAT*$(\mathcal{S})$* for $\mathcal{S} \in \{\mathcal{S}_{\mathsf{p}}, \mathcal{E}_{\mathsf{p}}, \mathcal{H}\}$ are given in [2]. The remaining cases are dealt with below.

LEMMA 3.4. *Let $Q = (I, D)$ be an instance of $\mathcal{A}^l$-SAT$(\mathcal{C}_{\mathsf{o}})$. Then Q is satisfiable if and only if D is satisfiable.*

Proof. Let $V = \{x_1, \dots, x_n\}$. If D is not satisfiable, then, obviously, the whole instance Q is not satisfiable. Suppose D is satisfiable, and $l(x_1) = a_1, \dots, l(x_n) = a_n$ is a solution of D. Then reorder variables in V so that $a_1 \leq \cdots \leq a_n$. Let $\epsilon = a_1/n$, and let, for $1 \leq i \leq n$, $f(x_i) = [\epsilon \cdot i, \epsilon \cdot i + a_i]$. It is easy to check that this f satisfies all constraints in Q. □

It follows that the problem $\mathcal{A}^l$-SAT$(\mathcal{C}_{\mathsf{o}})$ has exactly the same complexity as the problem HORNDLRSAT, and hence is tractable (see Theorem 2.5).

Algorithms for the remaining 6 subalgebras are given in Figure 1, and in the remainder of this subsection we prove that these algorithms are correct. (Checking that they are polynomial-time is straightforward and is left to the reader.)

Algorithms A_i, $1 \leq i \leq 4$, and Procedure P take an instance $Q = (I, D)$ over a set of variables V as input. We shall assume that D always contains all constraints of the form $l(v) > 0$, $v \in V$. We will also assume that I does not contain a constraint vrw, where $r = \emptyset$. This trivial necessary condition for satisfiability can obviously be checked in polynomial time.

The following lemma from [7] is crucial in our proofs of correctness.

LEMMA 3.5 (see [7]). *Let D be a satisfiable set of Horn DLRs, and let $x_1, \dots, x_n$ be the variables used in D. If $\tilde{D} = \{x_i \neq x_j \mid D \cup \{x_i \neq x_j\}$ is satisfiable$\}$, then $D \cup \tilde{D}$ is satisfiable.*

Using this lemma we can always divide the set of variables V into classes such that, in every model of an instance, variables from the same class must be assigned intervals of the same length while any variables from different classes can be assigned intervals of different lengths all at the same time.

LEMMA 3.6. *Algorithm A_1 correctly solves $\mathcal{A}^l$-SAT$(\mathcal{C}_{\mathsf{m}})$.*

Proof. Obviously, if A_1 rejects in line 1, then Q is not satisfiable.

Suppose A_1 rejects in line 3. Then G contains a simple cycle of odd length, $x_1, \dots, x_{2t+1}, x_1$. Then, in any model f of Q, all of the intervals $f(x_1), \dots, f(x_{2t+1})$ must have the same length, and hence, by definition of $\mathcal{C}_{\mathsf{m}}$, for all $1 \leq i \leq 2t$ we have $f(x_i)$ (mm^{-1}) $f(x_{i+1})$. These conditions imply that $f(x_1)$ $(\equiv \mathsf{pp}^{-1})$ $f(x_{2t+1})$. Therefore, it is impossible that $f(x_1)$ (mm^{-1}) $f(x_{2t+1})$, so Q is not satisfiable.

Suppose now that the algorithm accepts. We will show how to construct a model of Q. Note that in this case D is satisfiable. Let $V = \{x_1, \dots, x_n\}$, and let

```
Input: Instance $Q = (I, D)$ of $\mathcal{A}^l$-SAT$(\mathcal{S})$ with set of variables $V$
Algorithm $A_1$ for $\mathcal{S} = \mathcal{C}_{\mathsf{m}}$
  (1) If $D$ is not satisfiable, then reject;
  (2) Construct a graph $G = (V, E)$, where $(v, w) \in E$ if and only if
          – $D \cup \{l(v) \neq l(w)\}$ is not satisfiable, and
          – $vrw \in I$ for some $r$ such that $(\equiv) \not\subseteq r$;
  (3) If $G$ is 2-colorable, then accept; else reject.
Procedure $P$
  (1) Let $D' = D$;
  (2) For each $vrw \in I$ such that $r \subseteq (\mathsf{dsf})$ or $r \subseteq (\mathsf{d}^{-1}\mathsf{s}^{-1}\mathsf{f}^{-1})$,
        add the constraint $l(v) < l(w)$ or $l(v) > l(w)$, respectively, to $D'$;
  (3) For each $vrw \in I$ such that $(\equiv) \subseteq r \subseteq (\equiv \mathsf{dsf})$ or $(\equiv) \subseteq r \subseteq (\equiv \mathsf{d}^{-1}\mathsf{s}^{-1}\mathsf{f}^{-1})$,
        add the constraint $l(v) \leq l(w)$ or $l(v) \geq l(w)$, respectively, to $D'$;
  (4) If $D'$ is not satisfiable, then reject.
Algorithm $A_2$ for $\mathcal{S} \in \{\mathcal{D}_{\mathsf{s}}, \mathcal{D}_{\mathsf{f}}, \mathcal{D}_{\mathsf{d}}\}$
  (1) Call procedure $P$;
  (2) Accept.
Algorithm $A_3$ for $\mathcal{S} = \mathcal{D}'_{\mathsf{d}}$
  (1) Call procedure $P$;
  (2) Construct a graph $G = (V, E)$, where $(v, w) \in E$ if and only if
        $D' \cup \{l(v) \neq l(w)\}$ is not satisfiable;
  (3) Identify the connected components $S_1, \ldots, S_k$ of $G$;
  (4) For each $S_j$, let $I_j = I|_{S_j} = \{vrw \in I \mid v, w \in S_j\}$
                  and $I'_j = \{v\ r \cap (\equiv \mathsf{oo}^{-1})\ w \mid vrw \in I_j\}$;
  (5) Solve $I'_j$, $1 \leq j \leq k$, as instances of $\mathcal{A}$-SAT$(\mathcal{S}_{\mathsf{o}})$;
  (6) If every $I'_j$ is satisfiable, then accept; else reject.
Algorithm $A_4$ for $\mathcal{S} = \mathcal{D}''_{\mathsf{d}}$
  (1) Call procedure $P$;
  (2) Construct a graph $G = (V, E)$, where $(v, w) \in E$ if and only if
          – $D' \cup \{l(v) \neq l(w)\}$ is not satisfiable, and
          – $vrw \in I$ for some $r$ such that $(\equiv) \subseteq r \cap (\equiv \mathsf{pp}^{-1}\mathsf{oo}^{-1}\mathsf{mm}^{-1}) \subseteq (\equiv \mathsf{mm}^{-1})$;
  (3) Identify the connected components $S_1, \ldots, S_k$ of $G$;
  (4) For each $S_j$, let $I_j = I|_{S_j} = \{vrw \in I \mid v, w \in S_j\}$
                  and $I'_j = \{v\ r\ w \mid vrw \in I_j$ and $(\equiv) \not\subseteq r\}$;
  (5) If every $I'_j$ is empty, then accept; else reject.
```

FIG. 1. *Polynomial-time algorithms for the tractable cases of* $\mathcal{A}^l$-SAT.

$\tilde{D} = \{l(x_i) \neq l(x_j) \mid D \cup \{l(x_i) \neq l(x_j)\}$ is satisfiable$\}$. Then, by Lemma 3.5, $D \cup \tilde{D}$ is satisfiable. Let $l(x_1) = a_1, \ldots, l(x_n) = a_n$ be a solution of $D \cup \tilde{D}$. We know that G can be colored with two colors, say black and white. Now if x_i is black let $f(x_i) = [0, a_i]$; otherwise let $f(x_i) = [-a_i, 0]$. Obviously, this satisfies all constraints containing $(\equiv)$ because all constraints in I already allow $(\mathsf{mm}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})$. Suppose that $x_i r x_j \in I$ for some r such that $(\equiv) \not\subseteq r$. If $(x_i, x_j) \in E$ then x_i and x_j are of different colors, and we have $f(x_i)\ (\mathsf{mm}^{-1})\ f(x_j)$. Otherwise we know, by Lemma 3.5, that the lengths of $f(x_i)$ and $f(x_j)$ are different, which means that $f(x_i)\ (\mathsf{mm}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})\ f(x_j)$, as required. □

The next three algorithms use preprocessing Procedure P (see Figure 1). This procedure can obviously be performed in polynomial time. It is also easy to see that P does not change the set of solutions to an input.

Lemma 3.7. *Algorithm A_2 correctly solves problems $\mathcal{A}^l$-SAT$(\mathcal{S})$, where $\mathcal{S} \in \{\mathcal{D}_{\mathsf{s}}, \mathcal{D}_{\mathsf{f}}, \mathcal{D}_{\mathsf{d}}\}$.*

Proof. Obviously, if the algorithm rejects (in P), then the instance is not satisfiable.

Suppose now the algorithm accepts. Let $l(x_1) = a_1, \dots, l(x_n) = a_n$ be a solution of D' and order the variables in V so that $a_1 \leq \cdots \leq a_n$.

If $\mathcal{F} = \mathcal{D}_{\mathsf{s}}$, then let $f(x_i) = [0, a_i]$ for all i. The constraints added to D during preprocessing P ensure that this f is a model of Q. Similarly, if $\mathcal{F} = \mathcal{D}_{\mathsf{f}}$, then the mapping given by $f(x_i) = [-a_i, 0]$ for all i satisfies all constraints in Q. Finally, if $\mathcal{F} = \mathcal{D}_{\mathsf{d}}$, then the mapping $f(x_i) = [-a_i/2, a_i/2]$ for all i satisfies all constraints in Q. □

Lemma 3.8. *Algorithm A_3 correctly solves $\mathcal{A}^l$-SAT$(\mathcal{D}'_{\mathsf{d}})$.*

Proof. Suppose first that A_3 accepts on an input Q. We construct a model of Q as follows. Let $v_{j,l}$, $1 \leq l \leq |S_j|$, be the members of S_j, $1 \leq j \leq k$. Let $\tilde{D} = \{l(v) \neq l(w) \mid D' \cup \{l(v) \neq l(w)\}$ is satisfiable$\}$. By Lemma 3.5, $D' \cup \tilde{D}$ is satisfiable.

Let $l(v_{j,l}) = a_{j,l}$, where $1 \leq j \leq k$ and $1 \leq l \leq |S_j|$, be a solution of $D' \cup \tilde{D}$. Note that, for every $1 \leq j \leq |S_j|$, we have $a_{j,1} = \cdots = a_{j,|S_j|}$. Reorder the S_j's so that $a_{1,1} < a_{2,1} < \cdots < a_{k,1}$ holds. Let

$$\epsilon = \begin{cases} \min\{\frac{a_{i+1,1} - a_{i,1}}{3} \mid 1 \leq i < k\} & \text{if } k > 1, \\ 1 & \text{if } k = 1. \end{cases}$$

For all $1 \leq j \leq k$, let f_j be a model of I'_j (and then of I_j as well) and assume without loss of generality that the variables in I_j are ordered so that $f_j(v^-_{j,1}) \leq f_j(v^-_{j,2}) \leq \cdots \leq f_j(v^-_{j,|S_j|})$. By applying an appropriate translation and scaling, all models f_j can be chosen so that $0 < f_j(v^-_{j,1}) \leq \cdots \leq f_j(v^-_{j,|S_j|}) < \epsilon$.

Now we combine the models f_j of I_j into one model f of $Q = (I, D)$: let $f(v^-_{j,l}) = -j \cdot \epsilon + f_j(v^-_{j,l})$ and $f(v^+_{j,l}) = f(v^-_{j,l}) + a_{j,l}$ (see Figure 2).

We immediately see that f satisfies all length constraints and all constraints within each I_j. It is also easy to check that we have $f(v_{i,l})$ (d) $f(v_{i',l'})$ whenever $i < i'$. Due to the check in Procedure P, this satisfies all constraints between variables from different I_j's.

Assume now that algorithm A_3 rejects. We will show that Q is not satisfiable. The result holds trivially if A_3 rejects on line 1 (that is, in P). Assume to the contrary that some I'_j is not satisfiable but Q is satisfiable. Clearly, if Q is satisfiable, then the instance I_j has a model f with all intervals of the same length a. Then f is also a model of $I''_j = \{v\ r \cap (\equiv \mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1})\ w\ |vrw \in I_j\}$.

Reorder the variables in I_j so that $f(v^-_{j,1}) \leq f(v^-_{j,2}) \leq \cdots \leq f(v^-_{j,|S_j|})$, and suppose that $\{f(v^-_{j,l}) \mid 1 \leq l \leq |S_j|\} = \{b_1, \dots, b_t\}$, where $1 \leq t \leq |S_j|$ and $b_1 < \cdots < b_t$.

By definition of $\mathcal{D}'_{\mathsf{d}}$, every constraint allowing (pm) allows (o) as well. Therefore the function g defined by

$$g(v_{j,l}) = [a \cdot s/|S_j|, a \cdot s/|S_j| + a] \qquad \text{when } f(v^-_{j,l}) = b_s$$

is a model of I_j. Moreover, it is also a model of I'_j, a contradiction. □

Lemma 3.9. *Algorithm A_4 correctly solves $\mathcal{A}^l$-SAT$(\mathcal{D}''_{\mathsf{d}})$.*

Proof. If A_4 rejects in line 1 (that is, in P), then Q is obviously not satisfiable.

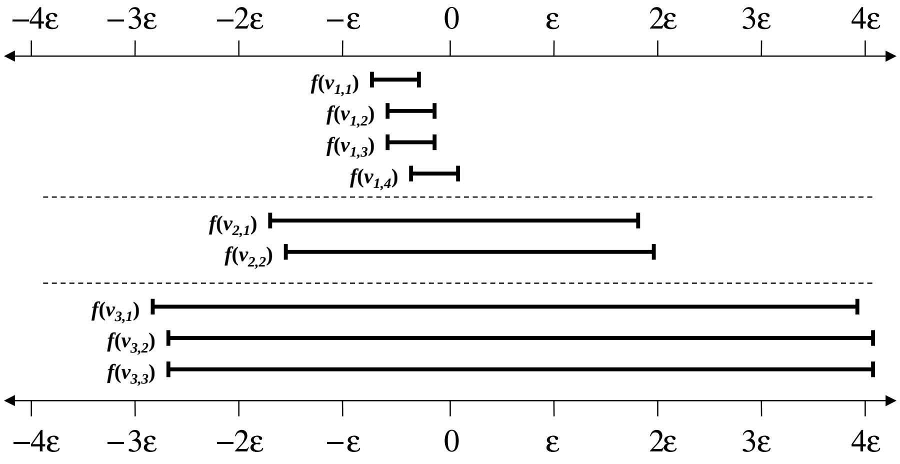


FIG. 2. *A combined model for an instance of* $\mathcal{A}^l$-SAT$(\mathcal{D}'_{\mathsf{d}})$ *(Lemma* 3.8*).*

Suppose A_4 rejects in line 6. It follows that there are variables $x_1, \dots, x_q \in V$ such that, in any model f of Q,

(i) intervals $f(x_1), \dots, f(x_q)$ have the same length, and

(ii) $f(x_i)$ $(\equiv \mathsf{mm}^{-1})$ $f(x_{i+1})$ for all $1 \leq i \leq q-1$, and

(iii) (by definition I'_j and $\mathcal{D}''_{\mathsf{d}}$) the intervals $f(x_1)$ and $f(x_q)$ are related by $(\mathsf{oo}^{-1}\mathsf{dd}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})$.

It is clear that these three conditions cannot be satisfied simultaneously. Therefore Q is not satisfiable.

Suppose that the algorithm accepts. We will show how to construct a model of Q. Let $v_{j,l}$, $1 \leq l \leq |S_j|$, be the members of S_j, $1 \leq j \leq k$. Let $\tilde{D} = \{l(v) \neq l(w) \mid D' \cup \{l(v) \neq l(w)\}$ is satisfiable$\}$. By Lemma 3.5, $D' \cup \tilde{D}$ is satisfiable.

Let $l(v_{j,l}) = a_{j,l}$, where $1 \leq j \leq k$ and $1 \leq l \leq |S_j|$, be a solution of $\tilde{D}$. Note that, for every $1 \leq j \leq |S_j|$, we have $a_{j,1} = \cdots = a_{j,|S_j|}$. Reorder the S_j's so that $a_{1,1} \leq a_{2,1} \leq \cdots \leq a_{k,1}$ holds (note that some of the $a_{j,1}$'s may coincide). Let $\{a_{1,1}, \dots, a_{k,1}\} = \{b_1, \dots, b_t\}$, where $b_1 < \cdots < b_t$, and let

$$\epsilon = \begin{cases} \min\{b_1, \frac{b_{i+1}-b_i}{3} \mid 1 \leq i < t\} & \text{if } t > 1, \\ 1 & \text{if } t = 1. \end{cases}$$

Further, let $f(v^-_{j,l}) = -s \cdot \epsilon + \frac{j}{|V|} \cdot \epsilon$, where s is such that $b_s = a_{j,l}$, and let $f(v^+_{j,l}) = f(v^-_{j,l}) + a_{j,l}$ (see Figure 3). We will show that f is a model of Q. By the choice of $a_{j,l}$, f satisfies all length constraints.

Suppose $v_{j,l}\ r\ v_{j',l'} \in I$ and check that $f(v_{j,l})\ r\ f(v_{j',l'})$.

Case 1. $j = j'$.

If the variables are from the same connected component of G, then we have that $(\equiv) \subseteq r$. Indeed, we have $f(v_{j,l})$ $(\equiv)$ $f(v_{j',l'})$ by the definition of f.

Case 2. $j \neq j'$, but $a_{j,l} = a_{j',l'}$.

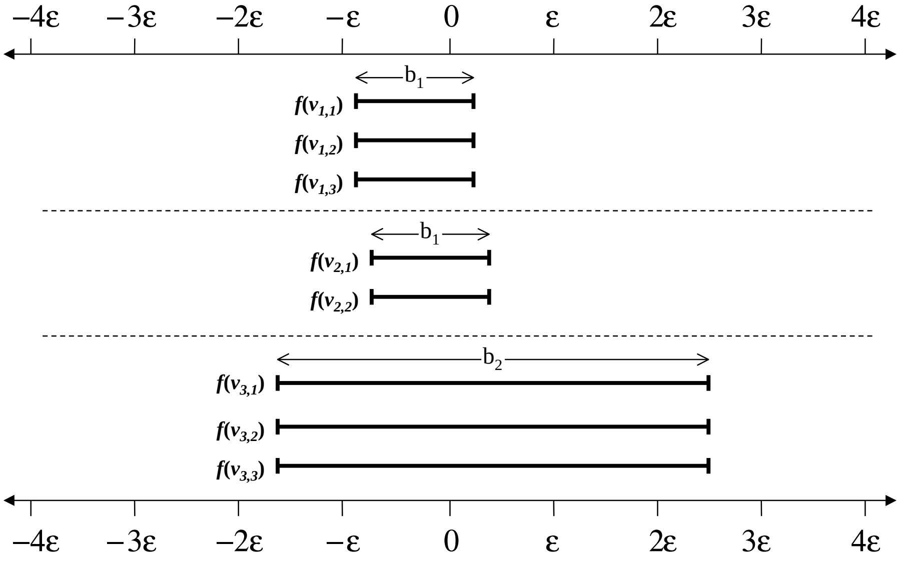


FIG. 3. *A combined model for an instance of* $\mathcal{A}^l$-SAT$(\mathcal{D}''_{\mathsf{d}})$ *(Lemma* 3.9*).*

By definition of G, we have either $r \cap (\mathsf{pp}^{-1}\mathsf{oo}^{-1}) \neq \emptyset$ or $(\equiv) \not\subseteq r$. In the former case we immediately get $(\mathsf{oo}^{-1}) \subseteq r$ by the definition of $\mathcal{D}''_{\mathsf{d}}$. Suppose that $(\equiv) \not\subseteq r$. Then $r \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}) = \emptyset$. Due to the check in P, the equality $a_{j,l} = a_{j',l'}$ is necessary. It follows from this fact and from the definition of $\mathcal{D}''_{\mathsf{d}}$ that we have $(\mathsf{oo}^{-1}) \subseteq r$ again. Indeed, it is easy to check that $f(v_{j,l})$ (o) $f(v_{j',l'})$ if $j < j'$, by the definition of f.

Case 3. $a_{j,l} \neq a_{j',l'}$.

Assume without loss of generality that $a_{j,l} < a_{j',l'}$. It follows from the definition of $\mathcal{D}''_{\mathsf{d}}$ that either we have $(\mathsf{dd}^{-1}) \subseteq r$ or (due to the check in P) $(\mathsf{d}) \subseteq r \subseteq (\mathsf{dsf})$. It is not hard to verify that, indeed, $f(v_{j,l})$ (d) $f(v_{j',l'})$, by the definition of f. □

3.2. NP-completeness results. First let us mention the obvious fact that, for any $\mathcal{F} \subseteq \mathcal{A}$, NP-completeness of $\mathcal{A}$-SAT$(\mathcal{F})$ implies NP-completeness of $\mathcal{A}^l$-SAT$(\mathcal{F})$.

LEMMA 3.10. *Suppose that $r_1, \ldots, r_n \in \mathcal{A}$ are relations such that the problem $\mathcal{A}$-*SAT*$(\{r_1, \ldots, r_n\})$ is NP-complete.*

1. *If, for every $1 \leq i \leq n$, $r'_i \in \{r_i, r_i \cup (\equiv)\}$, then $\mathcal{A}^l$-*SAT*$(\{r'_1, \ldots, r'_n\})$ is NP-complete.*
2. *If $\emptyset \neq r_1 \subseteq (\mathsf{pmo})$ and r'_1 satisfies $r_1 \subseteq r'_1 \subseteq r_1 \cup (\equiv \mathsf{dsf})$, then the problem $\mathcal{A}^l$-*SAT*$(\{r'_1, r_2, \ldots, r_n\})$ is NP-complete.*

Proof.

1. The proof is by polynomial-time reduction from $\mathcal{A}$-SAT($\{r_1, \dots, r_n\}$) to $\mathcal{A}^l$-SAT($\{r'_1, \dots, r'_n\}$). Let I be an instance of $\mathcal{A}$-SAT($\{r_1, \dots, r_n\}$) over a set V of variables. Construct an instance (I', D') of $\mathcal{A}^l$-SAT($\{r'_1, \dots, r'_n\}$) as follows:

(i) for every constraint urv in I such that $(\equiv) \subseteq r$ add urv to I';

(ii) for every constraint urv in I such that $(\equiv) \not\subseteq r$ add urv to I' and $l(u) \neq l(v)$ to D'.

Obviously, every solution to (I', D') is also a solution to I. Let f be a model of I, and let $\{x_1, \dots, x_m\}$ be the set of all endpoints of intervals $f(x)$, $x \in V$. We may without loss of generality assume that $0 < x_1 < \cdots < x_m$. Set $x'_1 = x_1$, $x'_2 = x_2$, and, for every $i > 2$, set $x'_i = 2x'_{i-1} + 1$. It is easy to check that the function f' such that $f'(v) = [x'_i, x'_j]$ if $f(v) = [x_i, x_j]$ is a model of (I', D').

2. Modify the previous construction as follows:

(i) for every constraint ur_1v in I add constraints ur'_1v to I' and $l(u) > l(v)$ to D';

(ii) for every constraint ur_iv, $i > 1$, in I add ur_iv to I'.

Every solution to (I', D') is also a solution to I because ur'_1v and $l(u) > l(v)$ imply ur_1v. Let f be a model of I, and let $\{x_1, \dots, x_m\}$ be the set of all endpoints of intervals $f(x)$ for some $x \in V$. We may without loss of generality assume that $x_1 < \cdots < x_m < 0$. Set $x'_m = x_m$, $x'_{m-1} = x_{m-1}$, and, for every $1 \leq i < m-1$, set $x'_i = 2x'_{i+1} - 1$. It is easy to check that the function f' such that $f'(v) = [x'_i, x'_j]$ if $f(v) = [x_i, x_j]$ is a model of (I', D'). □

Example 3.1. It follows from Theorem 2.2 that $\mathcal{A}$-SAT($\{(\mathsf{mm}^{-1})\}$) is NP-complete. Using Lemma 3.10(1) we conclude that $\mathcal{A}^l$-SAT($\{(\equiv \mathsf{mm}^{-1})\}$) is also NP-complete.

LEMMA 3.11. *$\mathcal{A}^l$-SAT($\mathcal{F}$) is NP-complete if $\mathcal{F}$ is* $\{(\mathsf{oo}^{-1}), (\mathsf{s})\}$, $\{(\mathsf{oss}^{-1}\mathsf{ff}^{-1})\}$, *or* $\{(\mathsf{sf}), (\mathsf{oo}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})\}$.

Proof. First let $\mathcal{F} = \{(\mathsf{oo}^{-1}), (\mathsf{s})\}$. The constraints

$$\{x(\mathsf{oo}^{-1})y,\ x(\mathsf{oo}^{-1})z,\ y(\mathsf{s})z;\ l(z) > l(x) + l(y)\}$$

are satisfiable if and only if $x(\mathsf{o})y$. Further, the constraints $\{x(\mathsf{o})z, z(\mathsf{o})y; l(z) > l(x) + l(y)\}$ are satisfiable if and only if $x(\mathsf{p})y$. It follows from Theorem 2.2 that $\mathcal{A}$-SAT($\{(\mathsf{oo}^{-1}), (\mathsf{p})\}$) is NP-complete. The above constructions show how to reduce $\mathcal{A}$-SAT($\{(\mathsf{oo}^{-1}), (\mathsf{p})\}$) to $\mathcal{A}^l$-SAT($\{(\mathsf{oo}^{-1}), (\mathsf{s})\}$) in polynomial time.

Now let $\mathcal{F} = \{(\mathsf{oss}^{-1}\mathsf{ff}^{-1})\}$. Note that in this case we can also make use of the relation $(\mathsf{ss}^{-1}\mathsf{ff}^{-1})$, which is equal to $(\mathsf{oss}^{-1}\mathsf{ff}^{-1})^*$.

We give a polynomial-time reduction from the NP-complete problem UNNEGATED ONE-IN-THREE 3SAT (Problem [LO4] in [11]) to $\mathcal{A}^l$-SAT($\{(\mathsf{oss}^{-1}\mathsf{ff}^{-1})\}$); let (X, C) be an arbitrary instance of UNNEGATED ONE-IN-THREE 3SAT. Consider the following set of constraints over the variables a, b, c, c':

$$\begin{array}{lll} a(\mathsf{oss}^{-1}\mathsf{ff}^{-1})b, & & l(a) = l(b) = 2, \\ c(\mathsf{ss}^{-1}\mathsf{ff}^{-1})a, & c(\mathsf{ss}^{-1}\mathsf{ff}^{-1})b, & l(c) = 1, \\ c'(\mathsf{ss}^{-1}\mathsf{ff}^{-1})a, & c'(\mathsf{ss}^{-1}\mathsf{ff}^{-1})b, & l(c') = 3. \end{array}$$

We impose the constraints $x(\mathsf{ss}^{-1}\mathsf{ff}^{-1})a$, $x(\mathsf{ss}^{-1}\mathsf{ff}^{-1})b$ on every $x \in X$ and note that this implies $l(x) \in \{1, 3\}$. To complete the reduction, we add the constraint $l(x) + l(y) + l(z) = 5$ for each $\{x, y, z\} \in C$. It is easy to show that the resulting set of constraints is satisfiable if and only if (X, C) has a solution.

Finally, let $\mathcal{F} = \{(\mathsf{sf}), (\mathsf{oo}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})\}$. The constraints $\{x(\mathsf{sf})z, y(\mathsf{sf})z; l(z) > l(x) + l(y)\}$ are satisfiable if and only if $x(\equiv \mathsf{pp}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})y$. Hence, we can obtain the relation $(\mathsf{ss}^{-1}\mathsf{ff}^{-1}) = (\mathsf{oo}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1}) \cap (\equiv \mathsf{pp}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})$. To show NP-completeness, use the same construction as above but replace $(\mathsf{oss}^{-1}\mathsf{ff}^{-1})$ with $(\mathsf{oo}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})$. □

LEMMA 3.12. *$\mathcal{A}^l$-SAT$(\{r\})$ is NP-complete whenever*

$$(\mathsf{mm}^{-1}) \subseteq r \subseteq (\mathsf{mm}^{-1}\mathsf{dd}^{-1}\mathsf{ss}^{-1}) \textit{ or } (\mathsf{mm}^{-1}) \subseteq r \subseteq (\mathsf{mm}^{-1}\mathsf{dd}^{-1}\mathsf{ff}^{-1}).$$

Proof. We consider only r with $(\mathsf{mm}^{-1}) \subseteq r \subseteq (\mathsf{mm}^{-1}\mathsf{dd}^{-1}\mathsf{ss}^{-1})$; the other case is dual. We may without loss of generality assume that $r = r^*$.

Case 1. $r = (\mathsf{mm}^{-1})$.

It follows from Theorem 2.2 that $\mathcal{A}$-SAT$(\{(\mathsf{mm}^{-1})\})$ is NP-complete.

Case 2. $r = (\mathsf{mm}^{-1}\mathsf{dd}^{-1})$.

The constraints

$$\begin{aligned} &x(\mathsf{dd}^{-1}\mathsf{mm}^{-1})y, \quad l(x) < l(y),\\ &a(\mathsf{dd}^{-1}\mathsf{mm}^{-1})x, \quad l(a) < l(x),\\ &a(\mathsf{dd}^{-1}\mathsf{mm}^{-1})y, \quad l(a) < l(y), \end{aligned}$$

are satisfiable if and only if $x(\mathsf{d})y$. Furthermore, the constraints

$$u(\mathsf{mm}^{-1}\mathsf{dd}^{-1})v, \quad x(\mathsf{d})u, \quad y(\mathsf{d})v, \quad l(u) = l(v),$$

are satisfiable if and only if $x(\mathsf{pp}^{-1})y$. It follows from Theorem 2.2 that the problem $\mathcal{A}$-SAT$(\{(\mathsf{d}), (\mathsf{pp}^{-1})\})$ is NP-complete. We have derived (d) and (pp^{-1}) from $(\mathsf{mm}^{-1}\mathsf{dd}^{-1})$, and hence $\mathcal{A}^l$-SAT$(\{(\mathsf{mm}^{-1}\mathsf{dd}^{-1})\})$ is also NP-complete.

Case 3. $r = (\mathsf{mm}^{-1}\mathsf{ss}^{-1})$.

The constraints

$$\begin{aligned} &a(\mathsf{mm}^{-1}\mathsf{ss}^{-1})x, \quad a(\mathsf{mm}^{-1}\mathsf{ss}^{-1})y, \quad l(x) > l(a),\\ &b(\mathsf{mm}^{-1}\mathsf{ss}^{-1})x, \quad b(\mathsf{mm}^{-1}\mathsf{ss}^{-1})y, \quad l(y) > l(b),\\ &x(\mathsf{mm}^{-1}\mathsf{ss}^{-1})y, \qquad\qquad\qquad\qquad l(x) = l(y), \end{aligned}$$

are satisfiable if and only if $a(\equiv\ \mathsf{ss}^{-1})b$, so we can derive the relation $(\equiv\ \mathsf{ss}^{-1})$. Furthermore, the constraints

$$\begin{aligned} &a(\equiv \mathsf{ss}^{-1})x, \qquad b(\equiv \mathsf{ss}^{-1})x, \quad l(x) > l(a),\\ &a(\equiv \mathsf{ss}^{-1})y, \qquad b(\equiv \mathsf{ss}^{-1})y, \quad l(y) > l(b),\\ &x(\mathsf{mm}^{-1}\mathsf{ss}^{-1})y, \qquad\qquad\qquad l(x) = l(y), \end{aligned}$$

are satisfiable if and only if $a(\mathsf{pp}^{-1})b$. Now NP-completeness follows from Theorem 2.2.

Case 4. $r = (\mathsf{mm}^{-1}\mathsf{dd}^{-1}\mathsf{ss}^{-1})$.

Replace $(\mathsf{mm}^{-1}\mathsf{ss}^{-1})$ with $(\mathsf{mm}^{-1}\mathsf{dd}^{-1}\mathsf{ss}^{-1})$ in the previous case. □

LEMMA 3.13. *$\mathcal{A}^l$-SAT$(\{r_1, r_2\})$ is NP-complete whenever $(\equiv) \not\subseteq r_2$ and*

$$r_1 \cap (\equiv \mathsf{pp}^{-1}\mathsf{oo}^{-1}\mathsf{mm}^{-1}) = (\mathsf{mm}^{-1}) \subsetneq r_2 \cap (\equiv \mathsf{pp}^{-1}\mathsf{oo}^{-1}\mathsf{mm}^{-1}).$$

Proof. Let us assume that all intervals have length one and prove that the problem $\mathcal{A}^l$-SAT$(\{r_1, r_2\})$ is NP-complete even under this assumption. This assumption reduces the number of cases to be considered because, in this case, we have $r_1 = (\mathsf{mm}^{-1})$ and $(\mathsf{mm}^{-1}) \subset r_2 \subseteq (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1})$. Moreover, we may without loss of generality assume that either $r_2^* = (\mathsf{mm}^{-1})$ or $r_2^* = r_2$.

Case 1. $\{(\mathsf{mm}^{-1}), (\mathsf{pmm}^{-1})\}$.

Let $G = (V, E)$ and $H = (V', E')$ denote two directed graphs. A homomorphism from G to H is a function $h : V \to V'$ such that $(v, w) \in E$ implies $(f(v), f(w)) \in E'$.

Let H be the graph $(V', E') = (\{0, 1, 2\}, \{(0, 1)(0, 2), (1, 0), (1, 2), (2, 1)\})$. Deciding whether there exists a homomorphism from an arbitrary graph to H is NP-complete, as follows from Theorem 4.4 in [27]. We denote this problem Graph Homomorphism(H).

We prove that $\{(\mathsf{mm}^{-1}), (\mathsf{pmm}^{-1})\}$ is NP-complete by a polynomial-time reduction from Graph Homomorphism(H). Arbitrarily choose a directed graph $G = (V, E)$.

The relations (pm) and (m) can be derived as follows. The constraints

$$\{x(\mathsf{mm}^{-1})x', y(\mathsf{pmm}^{-1})x, y(\mathsf{pmm}^{-1})x'\}$$

are satisfiable if and only if $y(\mathsf{pm})x$, and we have $(\mathsf{m}) = (\mathsf{pm}) \cap (\mathsf{mm}^{-1})$.

Introduce five fresh variables and the constraints $a(\mathsf{m})b(\mathsf{m})c(\mathsf{m})d(\mathsf{m})e$. For each node $v \in V$, add the constraints $a(\mathsf{pm})v(\mathsf{pm})e$. For each edge $(v, w) \in E$, add the constraint $v(\mathsf{pmm}^{-1})w$.

We show that the resulting set I of constraints are satisfiable if and only if there exists a homomorphism from G to H.

Only-if: Assume without loss of generality that f is a model of I such that $f(a) = [-1, 0]$. Construct a function $h : V \to V'$ as follows: $h(v) = \lfloor f(v^-) \rfloor$. To see that h is a homomorphism from G to H, arbitrarily choose an edge $(v, w) \in E$. We consider three cases:

(i) $h(v) = 0$. This implies that $0 \leq f(v^-) < 1$. Since $v(\mathsf{pmm}^{-1})w \in I$ and $f(w^+) \leq 3$, we know that $1 \leq f(w^-) \leq 2$ and $h(w) \in \{1, 2\}$. Hence, $(h(v), h(w)) \in E'$.

(ii) $h(v) = 1$. Either $0 \leq f(w^-) < 1$ (corresponding to $v(\mathsf{m}^{-1})w$) or $f(w^-) = 2$ (corresponding to $v(\mathsf{m})w$), so $h(w) \in \{0, 2\}$ and $(h(v), h(w)) \in E'$.

(iii) $h(v) = 2$. Then $f(w^-) = 1$ (corresponding to $v(\mathsf{m}^{-1})w$), $h(w) = 1$, and $(h(v), h(w)) \in E'$.

If: Assume $h : V \to V'$ is a homomorphism from G to H. Then f (as defined below) is a model of I:

$$f(a) = [-1, 0],\ f(b) = [0, 1],\ f(c) = [1, 2],\ f(d) = [2, 3],\ f(e) = [3, 4],$$

and for every $v \in V$ let $f(v) = [h(v), h(v) + 1]$.

Case 2. $\{(\mathsf{mm}^{-1}), (\mathsf{pp}^{-1}\mathsf{mm}^{-1})\}$.

The proof is by polynomial-time reduction from the NP-complete problem Graph 3-colorability (Problem [GT4] in [11]). Let $G = (V, E)$ be an arbitrary instance. Fix a fresh interval variable x. Introduce two interval variables v, v' for each $v \in V$ together with the constraints $v(\mathsf{mm}^{-1})v'(\mathsf{mm}^{-1})x$. Finally, add the constraint $v(\mathsf{pp}^{-1}\mathsf{mm}^{-1})w$ for every $(v, w) \in E$. It is easy to check that the resulting set of constraints is satisfiable if and only if G is 3-colorable. For example, if $f(x) = [3, 4]$, then constraints of the first type imply that $f(v) \in \{[1, 2], [3, 4], [5, 6]\}$ for any $v \in V$, while the constraints of the second type ensure that the values for "adjacent" variables are distinct.

Case 3. $\{(\mathsf{mm}^{-1}), (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1})\}$.

Use $(\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1})$ instead of $(\mathsf{pp}^{-1}\mathsf{mm}^{-1})$ in Case 2.

Case 4. $\{(\mathsf{mm}^{-1}), (\mathsf{mm}^{-1}\mathsf{oo}^{-1})\}$.

The proof is by polynomial-time reduction from the NP-complete problem BETWEENNESS[1] (Problem [MS1] in [11]), which is defined as follows:

INSTANCE: A finite set A, a collection T of ordered triples (a, b, c) of distinct elements from A.

QUESTION: Is there a total ordering $<$ on A such that for each $(a, b, c) \in T$ we have either $a < b < c$ or $c < b < a$?

Let (A, T) be an arbitrary instance of BETWEENNESS and note that the constraints $\{x(\mathsf{mm}^{-1})x',\ y(\mathsf{mm}^{-1}\mathsf{oo}^{-1})x, y(\mathsf{mm}^{-1}\mathsf{oo}^{-1})x'\}$ are satisfiable if and only if $x(\mathsf{oo}^{-1})y$. We construct an instance I over $\{(\mathsf{mm}^{-1}), (\mathsf{oo}^{-1})\}$ as follows:

(i) for each pair of distinct elements $a, b \in A$, add the constraint $a(\mathsf{oo}^{-1})b$ to I;

(ii) for each triple $(a, b, c) \in T$, introduce two fresh variables x, y and add the constraints $\{x(\mathsf{mm}^{-1})a,\ x(\mathsf{oo}^{-1})b,\ x(\mathsf{oo}^{-1})c,\ y(\mathsf{oo}^{-1})a,\ y(\mathsf{oo}^{-1})b,\ y(\mathsf{mm}^{-1})c\}$.

We will henceforth refer to the variables in I that correspond to the set A as "basic" variables and the other variables as "auxiliary" variables.

Assume that I has a model f. Then, due to the constraints added in step (i), the intervals $f(a), a \in A$, are pairwise distinct. Moreover, the relation (o) induces a total order on the set $\{f(a) \mid a \in A\}$. Suppose that there is a triple $(a, b, c) \in T$ such that the model f satisfies $f(b)\,(\mathsf{o})\,f(a)\,(\mathsf{o})\,f(c)$ and consider the constraints over the auxiliary variables x and y introduced in step (ii) for the triple (a, b, c). The variable x has to satisfy $x(\mathsf{mm}^{-1})a$, which implies that either $x(\mathsf{p})c$ or $x(\mathsf{p}^{-1})b$, a contradiction. We can analogously rule out all orderings of $f(a), f(b), f(c)$ except $f(a)\,(\mathsf{o})\,f(b)\,(\mathsf{o})\,f(c)$ and $f(c)\,(\mathsf{o})\,f(b)\,(\mathsf{o})\,f(a)$. Hence there is a solution to the instance (A, T): for all $a, b \in A$, set $a < b$ if and only if $f(a)\,(\mathsf{o})\,f(b)$.

Conversely, assume that there exists a total order $<$ on A that is a solution to the instance (A, T). We will show how to construct a model f of I. For all $a, b \in A$, set $f(a)\,(\mathsf{o})\,f(b)$ if and only if $a < b$. Clearly, this satisfies all constraints added in step (i). To show that there exists consistent values for all auxiliary variables, arbitrarily pick one triple $(a, b, c) \in T$ (corresponding to the auxiliary variables x and y) and assume without loss of generality that $a < b < c$. Let $a(\mathsf{m})x$, i.e., $f(x) = [f(a^-), f(a^-) + 1]$ and $y(\mathsf{m})c$; i.e., $f(y) = [f(c^-) - 1, f(c^-)]$. It is straightforward to verify that this construction satisfies all constraints.

Case 5. $\{(\mathsf{mm}^{-1}), (\mathsf{mm}^{-1}\mathsf{o})\}$.

The constraints $x(\mathsf{mm}^{-1})x',\ y(\mathsf{omm}^{-1})x,\ y(\mathsf{o}^{-1}\mathsf{mm}^{-1})x'$ are satisfiable if and only if $x(\mathsf{o}^{-1})y$. The constraints $x(\mathsf{o})x',\ y(\mathsf{o})x'$ are satisfiable if and only if $x(\equiv \mathsf{oo}^{-1})y$. The constraints $x(\equiv \mathsf{oo}^{-1})y,\ x'(\equiv \mathsf{oo}^{-1})x,\ x'(\mathsf{mm}^{-1})y$ are satisfiable if and only if $x(\mathsf{oo}^{-1})y$. Consequently, we can derive (mm^{-1}) and (oo^{-1}); continue as in Case 4.

Case 6. $\{(\mathsf{mm}^{-1}), (\mathsf{pmm}^{-1}\mathsf{o})\}$.

The constraints $x(\mathsf{mm}^{-1})x',\ y(\mathsf{pmm}^{-1}\mathsf{o})x,\ y(\mathsf{p}^{-1}\mathsf{mm}^{-1}\mathsf{o}^{-1})x'$ are satisfiable if and only if $x(\mathsf{o}^{-1})y$. Continue as in Case 5.

Case 7. $\{(\mathsf{mm}^{-1}), (\mathsf{pmm}^{-1}\mathsf{o}^{-1})\}$.

The constraints $x(\mathsf{mm}^{-1})x',\ y(\mathsf{pmm}^{-1}\mathsf{o}^{-1})x,\ y(\mathsf{pmm}^{-1}\mathsf{o}^{-1})x'$ are satisfiable if and only if $y(\mathsf{pm})x$. The relation $(\mathsf{m}) = (\mathsf{pm}) \cap (\mathsf{p}^{-1}\mathsf{mm}^{-1}\mathsf{o})$.

The constraints $a(\mathsf{m})b(\mathsf{m})c(\mathsf{m})d,\ a(\mathsf{pm})x,\ y(\mathsf{pm})d$ are satisfiable if and only if $x(\equiv \mathsf{mm}^{-1}\mathsf{oo}^{-1})y$. Hence, we can derive the relation $(\mathsf{pmm}^{-1}\mathsf{o}^{-1}) \cap (\equiv \mathsf{mm}^{-1}\mathsf{oo}^{-1}) = (\mathsf{mm}^{-1}\mathsf{o}^{-1})$, and NP-completeness follows from Case 5. □

3.3. Classification of complexity. The classification proof splits into 8 lemmas. In each lemma, it is proved that if a subalgebra $\mathcal{S}$ which is closed under deriva-

[1] This problem is also known as the TOTAL ORDERING PROBLEM [31].

tions with lengths satisfies a certain condition, then either $\mathcal{S}$ is contained in one of the 10 tractable subalgebras, or some lemma from section 3.2 can be applied to some subset of $\mathcal{S}$, or $\mathcal{S}$ satisfies the conditions of one of the previous lemmas. It is easy to verify that the assumptions of these 8 lemmas are exhaustive (note that, due to closedness under derivations with lengths, a subalgebra containing $r \cup (\equiv)$, where $r \subseteq (\mathsf{dsf})$, also contains r itself).

We can assume without loss of generality that each subalgebra $\mathcal{S}$ contains the total relation (the union of all basic relations), since we always allow pairs of variables to be unrelated. For each basic relation b of $\mathcal{A}$, we will write r_b to denote the least relation $r \in \mathcal{S}$ such that $(b) \subseteq r$, i.e., the intersection of all $r \in \mathcal{S}$ with this property. (Obviously, the relations r_b depend on $\mathcal{S}$; however, $\mathcal{S}$ will always be clear from the context.)

We use the relations of the form r_b in the algebraic proofs below to show that S is contained in one or another subalgebra. For example, suppose we know that the relation (p) is contained in r_{o}. Then any relation $r \in \mathcal{S}$ such that $(\mathsf{o}) \subseteq r$ satisfies also $(\mathsf{p}) \subseteq r$. To see this, note that if there is $r_1 \in \mathcal{S}$ such that $(\mathsf{o}) \subseteq r$, but $(\mathsf{p}) \not\subseteq r$, then $(\mathsf{o}) \subseteq r_1 \cap r_{\mathsf{o}}$ and $r_1 \cap r_{\mathsf{o}}$ is strictly contained in r_{o}, which contradicts the definition of r_{o}. By a similar argument, if we know that (p) is contained in all of r_{m}, r_{o}, r_{d}, and r_{s}, then we can conclude that, for every $r \in \mathcal{S}$, $(\mathsf{p}) \subseteq r$ whenever $r \cap (\mathsf{pmods}) \neq \emptyset$, which means that $\mathcal{S} \subseteq \mathcal{E}_{\mathsf{p}}$.

LEMMA 3.14. *Suppose $\mathcal{S}$ contains a nontrivial relation $r \subseteq (\equiv \mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1})$. Then either $\mathcal{S}$ is contained in one of $\mathcal{C}_{\mathsf{o}}$, $\mathcal{S}_{\mathsf{p}}$, $\mathcal{E}_{\mathsf{p}}$, and $\mathcal{H}$ or else $\mathcal{A}^l$-SAT$(\mathcal{S})$ is NP-complete.*

Proof.

Case 1. $r \subseteq (\equiv \mathsf{pp}^{-1}\mathsf{mm}^{-1})$.

If $\mathcal{S}$ is contained in one of $\mathcal{S}_{\mathsf{p}}$, $\mathcal{E}_{\mathsf{p}}$, and $\mathcal{H}$, then $\mathcal{A}^l$-SAT$(\mathcal{S})$ is tractable by Proposition 3.3. Otherwise let $\mathcal{S} = \{r_1, \dots, r_{n-1}\}$ and $r_n = r \setminus (\equiv)$ and apply Theorem 2.2 and Lemma 3.10(1) with $r_1, \dots, r_n$ to obtain NP-completeness of $\mathcal{A}^l$-SAT$(\mathcal{S})$.

Case 2. $r \cap (\mathsf{oo}^{-1}) = (\mathsf{o})$.

If $r^* \not\subseteq (\equiv)$, then the previous case applies. Assume that $r^* \subseteq (\equiv)$. If $r \not\subseteq (\equiv \mathsf{pmo})$, then using Lemma 3.10(1) one can show that $\mathcal{A}^l$-SAT$(\{r\})$ is NP-complete. If $(\mathsf{o}) \subseteq r \subseteq (\equiv \mathsf{pmo})$, then the constraints $\{xrz, zry; l(x)+l(y) < l(z)\}$ are satisfiable if and only if $x(\mathsf{p})y$. Therefore $(\mathsf{p}) \in \mathcal{S}$, and we go back to the first case.

Case 3. $(\mathsf{oo}^{-1}) \subseteq r$.

We may now assume that r is symmetric. We shall prove that either $\mathcal{S}$ is contained in one of $\mathcal{C}_{\mathsf{o}}$, $\mathcal{S}_{\mathsf{p}}$, and $\mathcal{E}_{\mathsf{p}}$ or else $\mathcal{A}^l$-SAT$(\mathcal{S})$ is NP-complete. Assume that $\mathcal{S} \not\subseteq \mathcal{C}_{\mathsf{o}}$; that is, there is $r' \in \mathcal{S}$ such that $(\mathsf{oo}^{-1}) \not\subseteq r'$. If $r \cap r' \not\subseteq (\equiv)$, then we obtain the required result by Cases 1 and 2. Therefore we may assume that $r \cap r'$ is either $\emptyset$ or $(\equiv)$ for every $r' \in \mathcal{S}$ such that $(\mathsf{oo}^{-1}) \not\subseteq r'$.

It now follows from Theorem 2.2 and Lemma 3.10(1) that if $\mathcal{S}$ is not contained in one of $\mathcal{S}_{\mathsf{o}}$, $\mathcal{E}_{\mathsf{o}}$, $\mathcal{S}_{\mathsf{p}}$, and $\mathcal{E}_{\mathsf{p}}$, then $\mathcal{A}^l$-SAT$(\mathcal{S})$ is NP-complete. If $\mathcal{S}$ is contained in $\mathcal{S}_{\mathsf{p}}$ or in $\mathcal{E}_{\mathsf{p}}$, then, by Proposition 3.3, $\mathcal{A}^l$-SAT$(\mathcal{S})$ is tractable. Suppose $\mathcal{S}$ is contained in $\mathcal{S}_{\mathsf{o}}$ or in $\mathcal{E}_{\mathsf{o}}$ but neither in $\mathcal{S}_{\mathsf{p}}$ nor in $\mathcal{E}_{\mathsf{p}}$. Then $\mathcal{S}$ contains a nontrivial symmetric relation r'' such that $(\mathsf{oo}^{-1}) \subseteq r'' \subseteq (\equiv \mathsf{mm}^{-1}\mathsf{oo}^{-1})$. Also, r' must be a nontrivial subrelation of $(\equiv \mathsf{ss}^{-1})$ or of $(\equiv \mathsf{ff}^{-1})$. We consider only the first case; the second is dual. Assume without loss of generality that $(\mathsf{s}) \subseteq r'$. Then the constraints $\{xr'y; l(x) < l(y)\}$ are satisfiable if and only if $x(\mathsf{s})y$. Therefore $(\mathsf{s}) \in \mathcal{S}$. Since $(r'' \circ (\mathsf{s}))^* = (\mathsf{oo}^{-1}) \in \mathcal{S}$, the problem $\mathcal{A}^l$-SAT$(\mathcal{S})$ is NP-complete by Lemma 3.11. ☐

LEMMA 3.15. *Suppose $\mathcal{S}$ contains a nontrivial relation r such that $r^* \subseteq (\equiv)$ and* *neither r nor r^{-1} is contained in $(\equiv \mathsf{dsf})$. Then either $\mathcal{S}$ is contained in one of $\mathcal{C}_{\mathsf{o}}$, $\mathcal{S}_{\mathsf{p}}$, $\mathcal{E}_{\mathsf{p}}$, and $\mathcal{H}$ or else $\mathcal{A}^l$-SAT$(\mathcal{S})$ is NP-complete.*

Proof. If neither $r \setminus (\equiv)$ nor $r^{-1} \setminus (\equiv)$ is contained in one of $(\mathsf{pmod}^{-1}\mathsf{sf}^{-1})$, $(\mathsf{pmod}^{-1}\mathsf{s}^{-1}\mathsf{f}^{-1})$, (pmodsf), or (pmodsf^{-1}), then $\mathcal{A}$-SAT$(\{r \setminus (\equiv)\})$ is NP-complete by Theorem 2.2, and we get the required result by Lemma 3.10(1).

Suppose now that $r \setminus (\equiv)$ is contained in one of the four relations above. Then (taking $r \circ r \circ r$ instead of r if needed) r can be chosen so that it satisfies one of the following conditions:

1. $r \subseteq (\equiv \mathsf{pmos})$;
2. $r \subseteq (\equiv \mathsf{pmof}^{-1})$;
3. $(\mathsf{pmosf}^{-1}) \subseteq r \subseteq (\equiv \mathsf{pmosf}^{-1})$;
4. $(\mathsf{pmods}) \subseteq r$;
5. $(\mathsf{pmod}^{-1}\mathsf{f}^{-1}) \subseteq r$.

Note that conditions 1 and 2 and conditions 4 and 5 are dual. Therefore it is sufficient to consider only conditions 1, 3, and 4.

Suppose condition 1 holds. Then, by assumption, $r \not\subseteq (\equiv \mathsf{s})$. Now it can be checked that the constraints $\{xrz, zry; l(x) > l(z)\}$ are satisfiable if and only if $xr'y$ for some nontrivial $r' \in \mathcal{A}$ such that $r' \subseteq (\mathsf{pmo})$. Then we apply Lemma 3.14.

Suppose condition 3 holds. Then the constraints $\{xrz, zry; l(x) < l(z), l(z) > l(y)\}$ are satisfiable if and only if $x(\mathsf{pmo})y$. Therefore we again apply Lemma 3.14.

Suppose condition 4 holds. If $(\equiv) \subseteq r$, then the constraints $\{xrz, zry; l(x) > l(z)\}$ are satisfiable if and only if $x(\equiv \mathsf{pmods})y$. Similarly, if $(\equiv) \not\subseteq r$, then the constraints $\{xrz, zry; l(x) > l(z)\}$ are satisfiable if and only if $x(\mathsf{pmods})y$. Therefore a relation $r_1 \in \mathcal{A}$ with $(\mathsf{pmods}) \subseteq r_1 \subseteq (\equiv \mathsf{pmods})$ belongs to $\mathcal{S}$.

If $\mathcal{S}$ contains a nontrivial relation $r_2 \subseteq r_1$ such that $(\mathsf{d}) \not\subseteq r_2$, then either r_2 satisfies condition 1 (and then we get the required result) or r_2 is one of (s), $(\equiv \mathsf{s})$. In the latter case the constraints $\{xr_2y; l(x) < l(y)\}$ are satisfiable if and only if $x(\mathsf{s})y$. So we have $(\mathsf{s}) \in \mathcal{S}$. Then the constraints $\{x(\mathsf{s})z, zr_1y; l(x) + l(y) = l(z)\}$ are satisfiable if and only if $x(\mathsf{p})y$. So we have $(\mathsf{p}) \in \mathcal{S}$, and we can apply Lemma 3.14.

From now on in this proof we assume that every nontrivial $r_2 \in \mathcal{S}$ such that $r_2 \subseteq r_1$ satisfies $(\mathsf{d}) \subseteq r_2$. It now follows that, for every $r \in \mathcal{S}$, $r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset$ implies $(\mathsf{d})^{\pm 1} \subseteq r$. In other words, we have $\mathcal{S} \subseteq \mathcal{E}_{\mathsf{d}}$.

If $(\mathsf{p}) \subseteq r_{\mathsf{d}}$, then, for every $r \in \mathcal{S}$, $r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset$ also implies $(\mathsf{pd})^{\pm 1} \subseteq r$, which means that $\mathcal{S} \subseteq \mathcal{E}_{\mathsf{p}}$, and we get the required result. If $(\mathsf{o}) \subseteq r_{\mathsf{d}}$, then, for every $r \in \mathcal{S}$, $r \cap (\mathsf{pmods})^{\pm 1} \neq \emptyset$ also implies $(\mathsf{od})^{\pm 1} \subseteq r$, and then it is easy to check that $\mathcal{S} \subseteq \mathcal{H}$.

Assume that $r_{\mathsf{d}} \subseteq (\equiv \mathsf{mds})$. If $(\mathsf{m}) \subseteq r_{\mathsf{d}}$, then it can be checked that the constraints $\{zr_{\mathsf{d}}x, zr_{\mathsf{d}}y; l(z) > l(y) > l(x)\}$ are satisfiable if and only if $x(\mathsf{s})y$. It is proved above that, in the presence of r_1 and (s), the required result holds.

Now we may assume that $(\mathsf{d}) \subseteq r_{\mathsf{d}} \subseteq (\equiv \mathsf{ds})$. Then r_{d} is either (d) or (ds) because $(\equiv)$ can be removed by adding the constraint $l(x) < l(y)$.

Assume now that $\mathcal{S} \not\subseteq \mathcal{H}$. It is easy to see that every relation in $\mathcal{S}$ satisfies condition 1 of $\mathcal{H}$. If there is $r_3 \in \mathcal{S}$ failing to satisfy condition 3 of $\mathcal{H}$, then $r_4 = r_3 \cap r_1$ satisfies $r_4 \subseteq (\mathsf{pmds})$ and $r_4 \cap (\mathsf{pm}) \neq \emptyset$. Then the constraints $\{xr_{\mathsf{d}}z, zr_4y; l(z) > l(y)\}$ are satisfiable if and only if $x(\mathsf{p})y$. Hence we have $(\mathsf{p}) \in \mathcal{S}$, and we can apply Lemma 3.14.

Finally, assume that every $r \in \mathcal{S}$ satisfies conditions 1 and 3 of $\mathcal{H}$, but some $r_5 \in \mathcal{S}$ fails to satisfy condition 2 of $\mathcal{H}$. We can assume that $r_5 \cap (\mathsf{ds}) \neq \emptyset$ and $r_5 \cap (\mathsf{d}^{-1}\mathsf{f}^{-1}) \neq \emptyset$, but $(\mathsf{o}) \not\subseteq r_5$. Let $\nu = (\equiv \mathsf{oo}^{-1}\mathsf{dd}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})$. Since $\nu = r_{\mathsf{d}}^{-1} \circ r_{\mathsf{d}}$

belongs to $\mathcal{S}$, we may assume that $r_5 \subseteq \nu$; otherwise replace r_5 by $r_5 \cap \nu$. Note that $(\mathsf{d}) \subseteq r_5$.

If $(\mathsf{o}^{-1}) \subseteq r_5$, then $(\mathsf{d}^{-1}) \subseteq r_5$ because $(\mathsf{d}) \subseteq r_{\mathsf{o}}$. Then the constraints in the set $\{xr_5^* y; l(x) \neq l(y)\}$ are satisfiable if and only if xr_6y, where $r_6 \in \mathcal{A}$ is a symmetric relation such that $(\mathsf{dd}^{-1}) \subseteq r_6 \subseteq (\mathsf{dd}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})$. We have $(\mathsf{pmods}) = r_1 \circ r_{\mathsf{d}} \in \mathcal{S}$. It follows from Theorem 2.2 that $\mathcal{A}$-SAT$(\{r_6, (\mathsf{pmo})\})$ is NP-complete. Then $\mathcal{A}^l$-SAT$(\mathcal{S})$ is NP-complete by Lemma 3.10(2).

Let $(\mathsf{o}^{-1}) \not\subseteq r_5$. If $(\mathsf{d}^{-1}) \subseteq r_5$, then the argument is as above. Otherwise we have $(\mathsf{df}^{-1}) \subseteq r_5 \subseteq (\equiv \mathsf{dsff}^{-1})$ (note that $(\mathsf{s}^{-1}) \not\subseteq r_5$ because $\mathcal{S} \subseteq \mathcal{E}_{\mathsf{d}}$). Then the constraints in the set $\{xr_5y; l(x) > l(y)\}$ are satisfiable if and only if $x(\mathsf{f}^{-1})y$. We may then assume that $(\mathsf{f}) \in \mathcal{S}$. It follows that the relations $(\mathsf{ods}) = (\mathsf{f}^{-1}) \circ r_{\mathsf{d}}$ and $(\equiv \mathsf{dff}^{-1}) = (\mathsf{f}) \circ r_5$ both belong to $\mathcal{S}$, and therefore $(\mathsf{d}) = r_{\mathsf{d}} \cap (\equiv \mathsf{dff}^{-1}) \in \mathcal{S}$. It follows from Theorem 2.2 that $\mathcal{A}$-SAT$(\{(\mathsf{o}), (\mathsf{d}), (\mathsf{f}), (\equiv \mathsf{dff}^{-1})\})$ is NP-complete. Since $(\mathsf{ods}) \in \mathcal{S}$, we conclude that $\mathcal{A}^l$-SAT$(\mathcal{S})$ is NP-complete by Lemma 3.10(2). □

LEMMA 3.16. *If $\mathcal{S}$ contains two nontrivial relations r_1 and r_2 such that $r_1 \cap r_2 \subseteq (\equiv)$ and $r_1, r_2 \subseteq (\equiv \mathsf{dsf})$, then either $\mathcal{S} \subseteq \mathcal{H}$ or else $\mathcal{A}^l$-SAT$(\mathcal{S})$ is NP-complete.*

Proof. We may assume that $(\equiv) \not\subseteq r_1, r_2$ because it can be removed by adding the constraint $l(x) < l(y)$. If $r_1 = (\mathsf{d})$ and $r_2 = (\mathsf{sf})$, then $\mathcal{A}$-SAT$(\{r_1, r_2\})$ is NP-complete by Theorem 2.2.

In all other cases $r_1 \circ r_2^{-1}$ (or its converse) satisfies the assumptions of Lemma 3.15 or Lemma 3.16. It remains to notice that $\{r_1, r_2\}$ is not contained in one of $\mathcal{C}_{\mathsf{o}}$, $\mathcal{S}_{\mathsf{p}}$, $\mathcal{E}_{\mathsf{p}}$. □

LEMMA 3.17. *If $\mathcal{S}$ contains two nontrivial symmetric relations r_1 and r_2 such that $r_1 \cap r_2 \subseteq (\equiv)$, then either $\mathcal{S}$ is contained in one of $\mathcal{S}_{\mathsf{p}}$, $\mathcal{E}_{\mathsf{p}}$, $\mathcal{H}$ or else $\mathcal{A}^l$-SAT$(\mathcal{S})$ is NP-complete.*

Proof. We may assume that r_1 and r_2 are minimal (with respect to inclusion) among nontrivial symmetric relations.

It follows from Theorem 2.2 that if none of r_1, r_2 is contained in one of $(\equiv \mathsf{ss}^{-1})$, $(\equiv \mathsf{ff}^{-1})$, then $\mathcal{A}$-SAT$(\mathcal{S})$ (and, consequently, $\mathcal{A}^l$-SAT$(\mathcal{S})$) is NP-complete.

We shall consider only the case $r_1 \subseteq (\equiv \mathsf{ss}^{-1})$; the case $r_1 \subseteq (\equiv \mathsf{ff}^{-1})$ is dual. Then we may assume that $(\mathsf{ss}^{-1}) \in \mathcal{S}$ and $(\mathsf{s}) \in \mathcal{S}$ because these constraints are equivalent to $\{xr_1y; l(x) \neq l(y)\}$ and $\{xr_1y; l(x) < l(y)\}$, respectively. If $r_2 \subseteq (\equiv \mathsf{dd}^{-1}\mathsf{ff}^{-1})$, then, by imposing the constraint $l(x) < l(y)$, we can obtain a nonempty subrelation of (df), and we can apply Lemma 3.16. We therefore may assume that $r_2 \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}) \neq \emptyset$. Now it follows from minimality of r_2 and from Theorem 2.2 that if $\mathcal{A}$-SAT$(\mathcal{S})$ is not NP-complete, then either $\mathcal{S} \subseteq \mathcal{H}$ or every relation $r \in \mathcal{S}$ such that $r^* \not\subseteq (\equiv \mathsf{ss}^{-1})$ satisfies $r_2 \subseteq r$.

It can be easily checked that if $(\mathsf{dd}^{-1}) \not\subseteq r_2$, then either $r_2 \subseteq (\equiv \mathsf{mm}^{-1})$ or $r_3 = ((\mathsf{s}) \circ r_2)^*$ is nonempty and satisfies $r_3 \subseteq (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1})$. In the former case $\mathcal{A}^l$-SAT$(\{r_2\})$ is NP-complete by Lemma 3.10(1). In the latter case we apply Lemma 3.14.

Further, let $(\mathsf{dd}^{-1}) \subseteq r_2$. Suppose some nontrivial relation $r_3 \in \mathcal{S}$ is strictly contained in r_2. Then, by the choice of r_2, we have $r_3^* \subseteq (\equiv)$, and, since $\mathcal{S} \not\subseteq \mathcal{C}_{\mathsf{o}}$, we can apply Lemma 3.15 or Lemma 3.16.

Now we may assume that, for every $r \in \mathcal{S}$ such that $r \cap r_2 \neq \emptyset$, we have $(\mathsf{dd}^{-1}) \subseteq r_2 \subseteq r$.

It can now be checked using Theorem 2.2 that if $\mathcal{A}$-SAT$(\mathcal{S})$ is not NP-complete, then $\mathcal{S}$ is contained in one of $\mathcal{S}_{\mathsf{p}}$, $\mathcal{S}_{\mathsf{d}}$, and $\mathcal{H}$. Suppose that $\mathcal{S} \not\subseteq \mathcal{S}_{\mathsf{p}}$ and $\mathcal{S} \not\subseteq \mathcal{H}$, since otherwise there is nothing to prove. Then $\mathcal{S} \subseteq \mathcal{S}_{\mathsf{d}}$, and for every relation $r \in \mathcal{S}$ such that $r \not\subseteq (\equiv \mathsf{ss}^{-1})$, we have $(\mathsf{dd}^{-1}) \subseteq r_2 \subseteq r$. If r_2 contains (pp^{-1}) or (oo^{-1}), then $\mathcal{S}$ is contained in $\mathcal{S}_{\mathsf{p}}$ or $\mathcal{H}$, which contradicts the assumptions just made. Otherwise we have $(\mathsf{mm}^{-1}\mathsf{dd}^{-1}) \subseteq r_2 \subseteq (\equiv \mathsf{mm}^{-1}\mathsf{dd}^{-1}\mathsf{ff}^{-1})$. Hence $((\mathsf{s}) \circ r_2)^* = (\mathsf{mm}^{-1}\mathsf{dd}^{-1}) \in \mathcal{S}$. By minimality, it follows that $r_2 = (\mathsf{mm}^{-1}\mathsf{dd}^{-1})$. Then $\mathcal{A}^l$-SAT$(\mathcal{S})$ is NP-complete by Lemma 3.12. ◻

LEMMA 3.18. *If* $(\mathsf{s}) \in \mathcal{S}$ *or* $(\mathsf{f}) \in \mathcal{S}$ *then either* $\mathcal{S}$ *is contained in one of the* 10 *subalgebras listed in Theorem* 3.1 *or else* $\mathcal{A}^l$-SAT$(\mathcal{S})$ *is NP-complete.*

Proof. We consider only the case $(\mathsf{s}) \in \mathcal{S}$; the other case is dual.

By Lemmas 3.15 and 3.16, we may assume that, for every nontrivial $r \in \mathcal{S}$ such that $r^* \subseteq (\equiv)$, we have either $(\mathsf{s}) \subseteq r \subseteq (\mathsf{dsf})$ or $(\mathsf{s}) \subseteq r^{-1} \subseteq (\mathsf{dsf})$. We may also assume that $(\mathsf{ss}^{-1}) \in \mathcal{S}$ because the constraints $\{x(\mathsf{s})z, z(\mathsf{s})y; l(x) \neq l(y)\}$ are satisfiable if and only if $x(\mathsf{ss}^{-1})y$.

Suppose that $\mathcal{S} \not\subseteq \mathcal{D}_{\mathsf{s}}$. Then there exists a relation $r_1 \in \mathcal{S}$ such that $r_1 \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}) \neq \emptyset$, but $(\equiv \mathsf{ss}^{-1}) \not\subseteq r_1$. If $(\equiv) \subseteq r_1$, then we can apply either Lemma 3.15 with r_1 or Lemma 3.17 with $\{r_1^*, (\mathsf{ss}^{-1})\}$. So we may now assume that $(\equiv) \not\subseteq r_1$. It can be checked that there is a nontrivial $r_2 \in \mathcal{A}$ such that $\{ur_1v, u(\mathsf{s})x, v(\mathsf{s})y; l(u) = l(v)\}$ is satisfiable if and only if xr_2y. Then $r_2 \in \mathcal{S}$. Moreover, we have $(\equiv \mathsf{ss}^{-1}) \cap r_2 = \emptyset$. If r_2 satisfies $r_2^* \subseteq (\equiv)$, then we apply Lemma 3.15 or Lemma 3.16. Otherwise $\{r_2^*, (\mathsf{ss}^{-1})\} \subseteq \mathcal{S}$, and we get the required result by Lemma 3.17. ◻

LEMMA 3.19. *If* $(\mathsf{sf}) \in \mathcal{S}$*, then either* $\mathcal{S} \subseteq \mathcal{D}_{\mathsf{s}}$ *or* $\mathcal{S} \subseteq \mathcal{D}_{\mathsf{f}}$ *or else* $\mathcal{A}^l$-SAT$(\mathcal{S})$ *is NP-complete.*

Proof. We have $(\mathsf{dsf}) = (\mathsf{sf}) \circ (\mathsf{sf}) \in \mathcal{S}$. We may assume that neither (s) nor (f) belong to $\mathcal{S}$; otherwise we obtain the result by Lemma 3.18, since, out of the 10 subalgebras, (sf) is contained only in $\mathcal{D}_{\mathsf{s}}$ and in $\mathcal{D}_{\mathsf{f}}$. It now follows that $(\mathsf{dsf})^{\pm 1} \cap r \neq \emptyset$ implies $(\mathsf{sf})^{\pm 1} \subseteq r$ for any $r \in \mathcal{S}$.

Suppose that $\mathcal{S}$ is not contained in $\mathcal{D}_{\mathsf{s}}$. Then there is $r_1 \in \mathcal{S}$ such that $(\equiv \mathsf{ss}^{-1}) \not\subseteq r_1$ and $r_1 \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}) \neq \emptyset$. Assume that $(\equiv) \subseteq r_1$. If $(\mathsf{ss}^{-1}) \cap r_1 = \emptyset$, then, by the previous paragraph, we have $r_1 \subseteq (\equiv \mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1})$, and we apply Lemma 3.14. Assume now that $(\mathsf{ss}^{-1}) \cap r_1 = (\mathsf{s})$. Then $r_1 \subseteq (\equiv \mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}\mathsf{dsf})$. Now we apply Lemma 3.14 if $r_1^* \not\subseteq (\equiv)$ and Lemma 3.15 otherwise.

Now assume that $(\equiv) \not\subseteq r_1$ and $r_1 \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}) \neq \emptyset$. If there is such an r_1 with the additional property that $r_1 \cap (\mathsf{oo}^{-1}) = \emptyset$, then the set of constraints $\{x(\mathsf{sf})u, ur_1v, y(\mathsf{sf})v; l(u) = l(v)\}$ is satisfiable if and only if $xr'y$, where $r' \in \mathcal{A}$ is some nontrivial relation such that $r' \subseteq (\mathsf{pp}^{-1}\mathsf{mm}^{-1})$. Then we can apply Lemma 3.14.

Suppose $r_1 \cap (\mathsf{oo}^{-1}) \neq \emptyset$. We have $(\equiv \mathsf{oo}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1}) = (\mathsf{s}^{-1}\mathsf{f}^{-1}) \circ (\mathsf{sf}) \in \mathcal{S}$. Consider $r_2 = r_1 \cap (\equiv \mathsf{oo}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})$. If $r_2 \subseteq (\mathsf{oo}^{-1})$, then $\mathcal{A}$-SAT$(\{(\mathsf{sf}), r_2\})$ is NP-complete by Theorem 2.2. Otherwise $(\mathsf{ss}^{-1}\mathsf{ff}^{-1}) \subseteq r_2$, and we have either $r_2 = (\mathsf{oss}^{-1}\mathsf{ff}^{-1})$ or $r_2 = (\mathsf{oo}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})$. In both cases $\mathcal{A}^l$-SAT$(\{(\mathsf{sf}), r_2\})$ is NP-complete by Lemma 3.11. ◻

LEMMA 3.20. *If there is* $r \in \mathcal{S}$ *such that* $(\mathsf{d}) \subseteq r \subseteq (\mathsf{dsf})$*, then either* $\mathcal{S}$ *is contained in one of the* 10 *subalgebras listed in Theorem* 3.1 *or else* $\mathcal{A}^l$-SAT$(\mathcal{S})$ *is NP-complete.*

Proof. Note that $\nu = (\equiv \mathsf{oo}^{-1}\mathsf{dd}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1}) = r^{-1} \circ r \in \mathcal{S}$.

We may assume that every $r \in \mathcal{S}$ such that $r^* \subseteq (\equiv)$ satisfies $(\mathsf{d}) \subseteq r \subseteq (\equiv \mathsf{dsf})$ or $(\mathsf{d}) \subseteq r^{-1} \subseteq (\equiv \mathsf{dsf})$; otherwise we apply Lemmas 3.15, 3.18, or 3.19. It follows, in particular, that no nontrivial subrelation of $(\equiv \mathsf{ss}^{-1}\mathsf{ff}^{-1})$ belongs to $\mathcal{S}$.

Suppose that there exists $r_1 \in \mathcal{S}$ such that $r_1 \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}) \neq \emptyset$, but $(\mathsf{dd}^{-1}) \not\subseteq r_1$. If $r_1^* \subseteq (\equiv)$, then we can apply Lemma 3.15 with r_1. Otherwise r_1^* is a symmetric nontrivial relation satisfying $(\mathsf{dd}^{-1}) \cap r_1^* = \emptyset$. If $r_1 \subseteq (\equiv \mathsf{pp}^{-1}\mathsf{mm}^{-1})$, then we can apply Lemma 3.14. Otherwise the relation $r_2 = \nu \cap r_1^* \in \mathcal{S}$ is nontrivial and satisfies $r_2 \subseteq (\equiv \mathsf{oo}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})$. We have $(\mathsf{oo}^{-1}) \subseteq r_2$ and $r_2 \cap r = \emptyset$, since no subrelation of (sf) belongs to $\mathcal{S}$. Now it is easy to verify that $\mathcal{A}$-SAT$(\{r_2, r\})$ is NP-complete, by Theorem 2.2.

From now on (in this proof) we may assume that, for every $r \in \mathcal{S}$, whenever $r \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}) \neq \emptyset$ we have $(\mathsf{dd}^{-1}) \subseteq r$. It now follows that condition 1 of $\mathcal{D}_{\mathsf{d}}$ and $\mathcal{D}'_{\mathsf{d}}$ is satisfied in $\mathcal{S}$.

Suppose there is $r_2 \in \mathcal{S}$ such that $r_2 \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}) \neq \emptyset$, but $r_2 \cap (\equiv \mathsf{oo}^{-1}) = \emptyset$. It is easy to check that there exists a nontrivial $r_3 \in \mathcal{A}$ with $r_3 \subseteq (\mathsf{pp}^{-1}\mathsf{mm}^{-1})$ such that $\{ur_2v, xru, yrv; l(u) = l(v)\}$ is satisfiable if and only if xr_3y (the relation r_3 depends on r and r_2). Then $r_3 \in \mathcal{S}$, and we can apply Lemma 3.14.

From now on (in this proof) we may also assume that, for every $r \in \mathcal{S}$, $r \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}) \neq \emptyset$ implies $r \cap (\equiv \mathsf{oo}^{-1}) \neq \emptyset$.

We know that $r_{\mathsf{o}} \subseteq \nu$. If $(\equiv) \subseteq r_{\mathsf{o}}$, then it is easy to check that $\mathcal{S} \subseteq \mathcal{D}_{\mathsf{d}}$.

Suppose $r_{\mathsf{o}} \cap (\equiv \mathsf{oo}^{-1}) = (\mathsf{o})$ and $\mathcal{S} \not\subseteq \mathcal{D}'_{\mathsf{d}}$. Then there is $r_4 \in \mathcal{S}$ such that $r_4 \cap (\mathsf{pm}) \neq \emptyset$, but $(\mathsf{o}) \not\subseteq r_4$. Then there exists a nontrivial $r_5 \in \mathcal{A}$ with $r_5 \subseteq (\mathsf{pm})$ such that the constraints

$$\{ur_{\mathsf{o}}z,\ zr_{\mathsf{o}}v,\ ur_4v,\ xru,\ yrv;\ l(u) = l(z) = l(v)\}$$

are satisfiable if and only if xr_5y. Then $r_5 \in \mathcal{S}$, and we can apply Lemma 3.14.

It remains to consider the case $r_{\mathsf{o}} \cap (\equiv \mathsf{oo}^{-1}) = (\mathsf{oo}^{-1})$. Then every $r_6 \in \mathcal{S}$ such that $r_6 \cap (\mathsf{oo}^{-1}) = \emptyset$, but $r_6 \cap (\mathsf{pp}^{-1}\mathsf{mm}^{-1}) \neq \emptyset$ satisfies $(\equiv) \subseteq r_6$.

If there is such r_6 with $r_6 \cap (\mathsf{pp}^{-1}) \neq \emptyset$, then there exists a nontrivial $r_7 \in \mathcal{A}$ with $r_7 \subseteq (\mathsf{pp}^{-1}\mathsf{mm}^{-1})$ such that the constraints

$$\{ur_{\mathsf{o}}z,\ zr_6v,\ ur_6v,\ xru,\ yrv;\ l(u) = l(z) = l(v)\}$$

are satisfiable if and only if xr_7y. Then $r_7 \in \mathcal{S}$, and we can apply Lemma 3.14.

Now we may assume that every $r \in \mathcal{S}$ with $r \cap (\mathsf{pp}^{-1}) \neq \emptyset$ also satisfies $(\mathsf{oo}^{-1}) \subseteq r$. Suppose $\mathcal{S} \not\subseteq \mathcal{D}'_{\mathsf{d}}$. Then there is $r_8 \in \mathcal{S}$ such that $(\mathsf{m}) \subseteq r_8$ and $r_8 \cap (\equiv \mathsf{pp}^{-1}\mathsf{oo}^{-1}) = (\equiv)$. Moreover, every $r \in \mathcal{S}$ such that $r \cap (\mathsf{mm}^{-1}) \neq \emptyset$ satisfies $(\equiv) \subseteq r$, since otherwise we can obtain a relation $r_9 (= r \cap r_8)$ such that $r_9 \cap (\equiv \mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1})$ is nonempty and is contained in (mm^{-1}), a contradiction.

Now either $\mathcal{S} \subseteq \mathcal{D}''_{\mathsf{d}}$ or else there is $r_{10} \in \mathcal{S}$ such that $(\mathsf{poo}^{-1}) \subseteq r_{10}$ and $(\equiv) \not\subseteq r_{10}$. In the latter case, again, there exists a relation $r_{11} \in \mathcal{A}$ with $r_{11} \subseteq (\mathsf{pp}^{-1}\mathsf{mm}^{-1})$ such that the constraints

$$\{ur_{10}z,\ zr_8^{-1}v,\ ur_8v,\ xru,\ yrv;\ l(u) = l(z) = l(v)\}$$

are satisfiable if and only if $xr_{11}y$. Then $r_{11} \in \mathcal{S}$ and we can apply Lemma 3.14. $\square$

LEMMA 3.21. *If there is a symmetric nontrivial relation $r' \in \mathcal{S}$ such that every nontrivial $r \in \mathcal{S}$ satisfies $r' \subseteq r$, then either $\mathcal{S}$ is contained in one of the* 10 *subalgebras listed in Theorem* 3.1 *or else $\mathcal{A}^l$-SAT$(\mathcal{S})$ is NP-complete.*

Proof. If r' contains (pp^{-1}), or (oo^{-1}), or $(\equiv \mathsf{dd}^{-1})$, or $(\equiv \mathsf{ss}^{-1})$, or $(\equiv \mathsf{ff}^{-1})$, then $\mathcal{S}$ is contained in $\mathcal{S}_{\mathsf{p}}$, or $\mathcal{C}_{\mathsf{o}}$, or $\mathcal{D}_{\mathsf{d}}$, or $\mathcal{D}_{\mathsf{s}}$, or $\mathcal{D}_{\mathsf{f}}$, respectively. If $r' \subseteq (\mathsf{dd}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1})$, then we can obtain an asymmetric relation in $\mathcal{S}$ which contradicts the assumption of this step. If $r' = (\equiv \mathsf{mm}^{-1})$, then $\mathcal{A}^l$-SAT$(\{r'\})$ is NP-complete by Example 3.1.

From now on (in this proof) we assume that all nontrivial $r \in \mathcal{S}$ satisfy the condition that $(\mathsf{mm}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1}) \subseteq r$; otherwise one of the earlier cases applies. If every nontrivial $r \in \mathcal{S}$ satisfies $(\equiv \mathsf{mm}^{-1}\mathsf{ss}^{-1}\mathsf{ff}^{-1}) \subseteq r$, then $\mathcal{S} \subseteq \mathcal{D}_{\mathsf{s}}$. Suppose that there is $r_1 \in \mathcal{S}$ such that $(\equiv) \not\subseteq r_1$. Then $r_1 \cap (\equiv \mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}) \subseteq (\mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1})$. If, for all $r \in \mathcal{S}$, $(\mathsf{pp}^{-1}) \subseteq r$ or, for all $r \in \mathcal{S}$, $(\mathsf{oo}^{-1}) \subseteq r$, then $\mathcal{S} \subseteq \mathcal{S}_{\mathsf{p}}$ or $\mathcal{S} \subseteq \mathcal{C}_{\mathsf{o}}$, respectively. Else, we can choose r_1 so that $r_1 \cap (\equiv \mathsf{pp}^{-1}\mathsf{mm}^{-1}\mathsf{oo}^{-1}) = (\mathsf{mm}^{-1})$. Now it is not hard to check that either $\mathcal{S} \subseteq \mathcal{C}_{\mathsf{m}}$ or else there is $r_2 \in \mathcal{S}$ such that the system $\{r_1, r_2\}$ satisfies the conditions of Lemma 3.13. $\square$

Classification is complete. Theorem 3.1 is proved.

4. Conclusion. In this paper we have given a complete classification of the complexity of interval satisfiability problems with very general length restrictions. Our main result, Theorem 3.1, determines the complexity of $\mathcal{A}^l$-SAT*($\mathcal{F}$)* for every possible subset $\mathcal{F} \subseteq \mathcal{A}$.

To conclude, we note that our NP-completeness proofs use only a very restricted subset of the allowable length constraints. In fact, we use constraints on lengths only of the following forms:

(i) comparing $l(x) + l(y)$ with $l(z)$,
(ii) comparing $l(x)$ and $l(y)$,
(iii) comparing $l(x)$ with a given number.

It follows that the NP-complete fragments of $\mathcal{A}^l$-SAT remain NP-complete even if we allow only these very limited forms of Horn DLRs to specify length constraints. This prompts us to make the following conjecture.

CONJECTURE 4.1. *All NP-complete cases of $\mathcal{A}^l$-*SAT *remain NP-complete if we allow fixing individual interval lengths as the only form of constraints on lengths.*

In fact, we suggest that an even stronger result may be true: it may be that in all cases where imposing restrictions on interval lengths causes intractability, simply requiring all intervals to have the same length will already be intractable.

PROBLEM 4.1. *Do all NP-complete cases of $\mathcal{A}^l$-*SAT *remain NP-complete if we search only for models with all intervals of the same length?*

REFERENCES

[1] J. ALLEN, *Maintaining knowledge about temporal intervals*, Comm. ACM, 26 (1983), pp. 832–843.

[2] O. ANGELSMARK AND P. JONSSON, *Some observations on durations, scheduling and Allen's algebra*, in Proceedings of the 6th Conference on Constraint Programming (CP'00), Lecture Notes in Comput. Sci. 1894, Springer-Verlag, Berlin, 2000, pp. 484–488.

[3] F. BARBER, *Reasoning on interval and point-based disjunctive metric constraints in temporal contexts*, J. Artificial Intelligence Res., 12 (2000), pp. 35–86.

[4] A. CARRANO, *Establishing the order of human chromosome-specific DNA fragments*, in Biotechnology and the Human Genome: Innovations and impact, Basic Life Sciences 46, A. Woodhead and B. Barnhart, eds., Plenum Press, London, New York, 1988, pp. 37–50.

[5] D. CORNEIL, H. KIM, S. NATARAJAN, S. OLARIU, AND A. SPRAGUE, *Simple linear time recognition of unit interval graphs*, Inform. Process. Lett., 55 (1995), pp. 99–104.

[6] N. CREIGNOU, S. KHANNA, AND M. SUDAN, *Complexity Classification of Boolean Constraint Satisfaction Problems*, SIAM Monogr. Discrete Math. Appl. 7, SIAM, Philadelphia, 2001.

[7] T. DRAKENGREN AND P. JONSSON, *Eight maximal tractable subclasses of Allen's algebra with metric time*, J. Artificial Intelligence Res., 7 (1997), pp. 25–45.

[8] T. DRAKENGREN AND P. JONSSON, *A complete classification of tractability in Allen's algebra relative to subsets of basic relations*, Artificial Intelligence, 106 (1998), pp. 205–219.

[9] S. FORTUNE, J. HOPCROFT, AND J. WYLLIE, *The directed subgraph homeomorphism problem*, Theoret. Comput. Sci., 10 (1980), pp. 111–121.

[10] C. GABOR, K. SUPOWIT, AND W. HSU, *Recognizing circle graphs in polynomial time*, J. Assoc. Comput. Mach., 36 (1989), pp. 435–473.

[11] M. GAREY AND D. JOHNSON, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, New York, 1979.

[12] P. Goldberg, M. Golumbic, H. Kaplan, and R. Shamir, *Four strikes against physical mapping of DNA*, Journal of Computational Biology, 2 (1995), pp. 139–152.

[13] M. Golumbic, H. Kaplan, and R. Shamir, *On the complexity of DNA physical mapping*, Adv. Appl. Math., 15 (1994), pp. 251–261.

[14] M. Golumbic, H. Kaplan, and R. Shamir, *Graph sandwich problems*, J. Algorithms, 19 (1995), pp. 449–473.

[15] M. Golumbic and E. Scheinerman, *Containment graphs, posets and related classes of graphs*, in Proceedings of the 3rd International Conference on Combinatorial Mathematics, Ann. New York Acad. Sci. 555, New York Acad. Sci., New York, 1989, pp. 192–204.

[16] M. Golumbic and R. Shamir, *Complexity and algorithms for reasoning about time: A graph-theoretic approach*, J. Assoc. Comput. Mach., 40 (1993), pp. 1108–1133.

[17] P. Hell and J. Nešetřil, *On the complexity of H-coloring*, J. Combin. Theory Ser. B, 48 (1990), pp. 92–110.

[18] W.-L. Hsu, *A simple test for interval graphs*, in Proceedings of the 18th International Workshop on Graph-Theoretic Concepts in Computer Science, Lecture Notes in Comput. Sci. 657, W.-L. Hsu and R. Lee, eds., Springer-Verlag, Berlin, 1992, pp. 11–16.

[19] P. Jeavons, *On the algebraic structure of combinatorial problems*, Theoret. Comput. Sci., 200 (1998), pp. 185–204.

[20] P. Jonsson and C. Bäckström, *A unifying approach to temporal constraint reasoning*, Artificial Intelligence, 102 (1998), pp. 143–155.

[21] R. Karp, *Mapping the genome: Some combinatorial problems arising in molecular biology*, in Proceedings of the 25th Symposium on the Theory of Computing (STOC'93), San Diego, CA, 1993, ACM, New York, 1993, pp. 278–285.

[22] H. Kautz and P. Ladkin, *Integrating metric and qualitative temporal reasoning*, in Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI'91), Anaheim, CA, 1991, pp. 241–246.

[23] L. Kirousis and P. Kolaitis, *The complexity of minimal satisfiability problems*, in Proceedings of the 18th Symposium on Theoretical Aspects of Computer Science (STACS 2001), Lecture Notes in Comput. Sci. 2010, Springer-Verlag, Berlin, 2001, pp. 407–418.

[24] M. Koubarakis, *Tractable disjunctions of linear constraints: Basic results and applications to temporal reasoning*, Theoret. Comput. Sci., 266 (2001), pp. 311–339.

[25] A. Krokhin, P. Jeavons, and P. Jonsson, *Reasoning about temporal relations: The tractable subalgebras of Allen's interval algebra*, J. ACM, 50 (2003), pp. 591–640.

[26] R. Ladner, *On the structure of polynomial time reducibility*, J. Assoc. Comput. Mach., 22 (1975), pp. 155–171.

[27] H. Maurer, J. Sudborough, and E. Welzl, *On the complexity of the general coloring problem*, Inform. and Control, 51 (1981), pp. 128–145.

[28] I. Meiri, *Combining qualitative and quantitative constraints in temporal reasoning*, Artificial Intelligence, 87 (1996), pp. 343–385.

[29] U. Montanari, *Networks of constraints: Fundamental properties and applications to picture processing*, Information Sci., 7 (1974), pp. 95–132.

[30] B. Nebel and H.-J. Bürckert, *Reasoning about temporal relations: A maximal tractable subclass of Allen's interval algebra*, J. Assoc. Comput. Mach., 42 (1995), pp. 43–66.

[31] J. Opatrny, *Total ordering problem*, SIAM J. Comput., 8 (1979), pp. 111–114.

[32] I. Pe'er and R. Shamir, *Realizing interval graphs with size and distance constraints*, SIAM J. Discrete Math., 10 (1997), pp. 662–687.

[33] I. Pe'er and R. Shamir, *Satisfiability problems on intervals and unit intervals*, Theoret. Comput. Sci., 175 (1997), pp. 349–372.

[34] T. Schaefer, *The complexity of satisfiability problems*, in Proceedings of the 10th Symposium on the Theory of Computing (STOC'78), San Diego, CA, 1978, ACM, New York, 1978, pp. 216–226.

[35] A. Webber, *Proof of the interval satisfiability conjecture*, Ann. Math. Artificial Intelligence, 15 (1995), pp. 231–238.

SIAM J. DISCRETE MATH.
Vol. 17, No. 3, pp. 478–519

THE DEGREE-DIAMETER PROBLEM FOR SEVERAL VARIETIES OF CAYLEY GRAPHS I: THE ABELIAN CASE[*]

RANDALL DOUGHERTY[†] AND VANCE FABER[‡]

Abstract. We consider the degree-diameter problem for Cayley graphs of Abelian groups (Abelian graphs) for both directed graphs and undirected graphs. The problem is closely related to that of finding efficient lattice coverings of Euclidean space by shapes such as octahedra and tetrahedra; we exploit this relationship in both directions. For two generators (dimensions), these methods yield optimal Abelian graphs with a given diameter k. (The results in two dimensions are not new; they are given in the literature of distributed loop networks.) We find an undirected Abelian graph with three generators and a given diameter k, which we conjecture to be as large as possible; for the directed case, we obtain partial results. These results are connected to efficient lattice coverings of $\mathbf{R}^3$ by octahedra or by tetrahedra; computations on Cayley graphs lead us to such lattice coverings, which we conjecture to be optimal. (The problem of finding such optimal coverings can be reduced to a finite number of nonlinear optimization problems.) We discuss the asymptotic behavior of the Abelian degree-diameter problem for large numbers of generators. The graphs obtained here are substantially better than traditional toroidal meshes, but, in the simpler undirected cases, retain certain desirable features such as good routing algorithms, easy constructibility, and the ability to host mesh-connected numerical algorithms without any increase in communication times.

Key words. degree-diameter, distributed loop network, lattice covering

AMS subject classifications. 05C12, 05C25, 11H31, 52C17, 68R10, 90B10

DOI. 10.1137/S0895480100372899

1. Introduction.

The degree-diameter problem for graphs can be stated as the following question: What is the largest number of vertices that a graph (undirected or directed) can have if one is given upper bounds on the degree of each vertex and on the diameter of the graph (the maximum path-distance from any vertex to any other)? One application of such graphs is in the design of interconnection networks for parallel processors, where one wants to have a large number of processors without requiring a large number of wires at a single processor or incurring long delays in communication from one processor to another. For more information on the (undirected) degree-diameter problem, see Dinneen and Hafner [10]; up-to-date results can be found online at http://www-mat.upc.es/grup_de_grafs/grafs/taula_delta_d.html.

A desirable extra property of such networks is that they appear identical from any processor. This means that the graphs one uses should be vertex-transitive; i.e., for any two vertices x and y, there is an automorphism of the graph which maps x to y. Here we will restrict ourselves to a special class of vertex-transitive graphs called Cayley graphs. A Cayley graph is specified by a group and a set of generators for this group; the vertices of the graph are the elements of the group, and the graph has an edge from x to y if and only if there is a generator g such that $y = xg$. (It can be shown that every vertex-transitive graph is isomorphic to a generalized form of Cayley graph called a Cayley coset graph [20]. In this paper, though, we will look only at Cayley graphs. For Abelian groups, this is no loss of generality since every Cayley

[*] Received by the editors May 26, 2000; accepted for publication (in revised form) September 9, 2003; published electronically March 30, 2004.
http://www.siam.org/journals/sidma/17-3/37289.html

[†] IDA Center for Communications Research, 4320 Westerra Ct., San Diego, CA 92121 (rdough@ccrwest.org). This author was supported by NSF grant DMS-9158092 and by a fellowship from the Sloan Foundation.

[‡] Atto Bioscience, 15010 Broschart Road, Rockville, MD 20850 (vfaber@atto.com).

coset graph of an Abelian group is isomorphic to a Cayley graph of an Abelian group.) In a directed Cayley graph from a group on d generators, every vertex has in-degree and out-degree d; if d generators are used to form an undirected Cayley graph, then the degree of each vertex is the number of generators of order 2 plus twice the number of generators of order greater than 2 (unless there are redundant generators). Thus we will usually discuss Cayley graphs on a given number of generators rather than of a given degree; the cases where some generators have order 2, and hence contribute only 1 rather than 2 to the degree of an undirected Cayley graph, will be handled separately.

The most straightforward approach to finding large Cayley graphs of small diameter on a given number d of generators is to examine various groups, look at some or all possible sets of d generators for such a group, and check whether each such set in fact generates the group and, if so, determine the diameter of the graph. But this can be a very large task even for relatively small groups and generating sets. In this paper, we will use a different approach which facilitates studying many groups and generating sets at once; it yields provably optimal results for some families of groups and good lower and upper bounds for others.

In its most general form, the idea is as follows. Let F be the free (universal) group on d generators. Then, for any group G and any set of d generators for G, there is a homomorphism $\pi : F \to G$ which maps the canonical generators for F to the given generators for G; clearly, π is surjective. Let N be the kernel of π. Then N is a normal subgroup of F, and $|F : N| = |G|$; in fact, G is isomorphic to F/N, and the Cayley graph of G with the given generators is isomorphic to the Cayley graph of F/N with the canonical generators for F. Let S be the set of elements of F which can be expressed as a word of length at most k in the generators. (In undirected Cayley graphs, such words may use inverse generators g^{-1} as well as generators; for the directed case, only words using generators, not inverse generators, are allowed.)

Proposition 1.1. *The Cayley graph for G on the given generators has diameter at most k if and only if $SN = F$.*

Proof. First, suppose $SN = F$. Let a be an arbitrary element of G; then $a = \pi(x)$ for some x, and x can be written in the form wy with $w \in S$ and $y \in N$. Hence, $a = \pi(wy) = \pi(w)\pi(y) = \pi(w)$. Now w can be written as a word of length at most k in the generators of F, so $a = \pi(w)$ can be written as the same word in the corresponding generators of G. Since a was arbitrary, the Cayley graph has diameter at most k.

Now suppose that the Cayley graph has diameter at most k. Let x be any element of F; then $\pi(x)$ can be written as a word w' of length at most k in the generators of G. Let w be the corresponding word in the generators of F; then $\pi(w) = w' = \pi(x)$, so $\pi(w^{-1}x)$ is the identity of G, and we have $w^{-1}x \in N$. Hence, $x = w(w^{-1}x) \in SN$, as desired. □

So finding a Cayley graph on d generators with diameter k whose size is as large as possible is equivalent to finding a normal subgroup N of F such that $SN = F$ and $|F : N|$ is as large as possible. Of course, we immediately get the upper bound $|F : N| \leq |S|$, but this is probably not attainable.

Unfortunately, the collection of normal subgroups of F is so large and varied as to be unmanageable. So what we will do instead is restrict ourselves to certain families (usually varieties) of groups; this allows us to replace F with a free group for the family in question, which may be much easier to work with. For instance, if we consider only the Cayley graphs of Abelian groups, then we can replace F with the free *Abelian* group on d generators, which is simply $\mathbf{Z}^d$; the normal subgroups of $\mathbf{Z}^d$ are well understood and relatively easy to work with. It turns out that this reduces the degree-diameter problem for Abelian Cayley graphs to interesting problems about lattice coverings of Euclidean space by various shapes. Some of these problems can be solved completely, giving optimal Abelian Cayley graphs; others are still open.

A simple path-counting argument gives upper bounds for the size n of a Cayley graph with d generators and diameter limit k: in the directed case,

$$n \leq 1 + d + d^2 + \cdots + d^k = \frac{d^{k+1} - 1}{d - 1},$$

and in the undirected case,

$$n \leq 1 + 2d + 2d(2d-1) + \cdots + 2d(2d-1)^{k-1} = \frac{d(2d-1)^k - 1}{d - 1}.$$

(The formulas for $d = 1$ are $k + 1$ and $2k + 1$.) These limits are well known and actually apply to the degree-diameter problem for arbitrary graphs; they are known as the Moore bounds. For $d = 1$, these limits are attained by simple cycle graphs, which are Cayley graphs of cyclic groups.

In most cases, we will find that the attainable values for n using Cayley graphs in the families we consider do not approach these upper bounds; the equations defining the families force many paths to be redundant. But the extra structure provided by the groups may provide compensating advantages in parallel computers, such as good routing algorithms, easy constructibility, and the ability to map common problems onto the architecture. In particular, many of the Cayley graphs of Abelian groups that we discuss in this paper are multidimensional rectangular meshes with additional connections at the boundary. Thus, mesh calculations with natural boundary conditions are trivially mapped into these graphs, while the extra connections are utilized only when global communications are being carried out. In addition, the mesh nature of these graphs allows the physical construction of the network to be carried out with relatively short wires. This will be discussed further below.

In a separate paper, we will examine other varieties of groups for which similar analyses of Cayley graphs are feasible.

2. Abelian groups. In the rest of this paper, we will examine the Cayley graphs arising from Abelian groups. Toroidal meshes and hypercubes are examples of such graphs.

The degree-diameter problem for Abelian Cayley graphs has been considered by others. In particular, Annexstein and Baumslag [2] show that the number of generators d, diameter k, and size n of a directed Abelian Cayley graph satisfy

$$k \geq \Omega(n^{1/d});$$

in fact, if $d \leq n^{1/d}$, then

$$k \geq \Omega(dn^{1/d}).$$

They also discuss similar results for Cayley graphs of nilpotent groups.

In addition, Chung [6] has constructed directed Abelian Cayley graphs G with $n = p^t - 1$, $d = p$, and

$$k \leq \left\lceil 2t + \frac{4t \log t}{\log p - 2\log(t-1)} \right\rceil$$

for any positive integer $t < \sqrt{p} + 1$, where p is a prime. (Note that Chung's examples have diameters which are small compared to the number of generators; in contrast, we will concentrate here on graphs with a small fixed number of generators and relatively large diameters.) Chung's methods involve estimates of the second eigenvalue of the Laplacian of the adjacency matrix of G. For more about estimating the diameter of general graphs from knowledge of this eigenvalue, see Chung, Faber, and Manteuffel [7]. This eigenvalue is also connected to the sphere *packing* problem for real lattices; see Urakawa [23].

The more specific case of Cayley graphs of *cyclic* groups has been studied more extensively; such graphs are usually referred to by some variant of the phrase "loop networks." The survey paper of Bernard, Comellas, and Hsu [3] is an excellent guide to the literature in this area.

We start by taking care of some generalities and notational matters. We will use the symbol + for the group operation(s). Let $\mathbf{Z}_m$ be the cyclic group of order m (for definiteness, the set $\{0, 1, \dots, m-1\}$ with the operation of addition modulo m). Each of the groups $\mathbf{Z}_m$ $(m = 1, 2, 3, \dots)$ and $\mathbf{Z}$ (the infinite cyclic group) has a canonical generator 1, but each also has other sets of generators, some of which will be important later.

When one has groups $G_1, G_2, \dots, G_l$ and a set of generators for each, then one can get a set of generators for the product $G_1 \times G_2 \times \cdots \times G_l$ by putting together the given generator sets. More precisely, for each $i \leq l$ and each generator g of G_i, let $\mathbf{e}_i(g)$ be the element of the product group which has the identity element of G_j as its jth coordinate for all $j \leq l$ except i; the ith coordinate is g. Then the set of all elements $\mathbf{e}_i(g)$ is a natural generating set for the product group. (The resulting diameter for the product group is the sum of the diameters of the groups G_i.) In the case where the groups G_i are cyclic groups with canonical single generators, we write simply $\mathbf{e}_i$ for the l-tuple with 1 at the ith coordinate and 0 elsewhere.

A two-dimensional toroidal mesh is simply the Cayley graph of the group $\mathbf{Z}_m \times \mathbf{Z}_n$ with the canonical generators $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$; higher-dimensional meshes are obtained from longer products. For the two-dimensional case, the number of vertices is mn, the degree is 2 in the directed case and 4 in the undirected case (assuming $m, n \geq 3$), and the diameter is $m + n - 2$ in the directed case and $\lfloor m/2 \rfloor + \lfloor n/2 \rfloor$ in the undirected case. The calculations in three or more dimensions are analogous.

The d-dimensional hypercube is the Cayley graph of the group $\mathbf{Z}_2^d$ with the canonical generators; since we get bidirectional links in any case, we may as well discuss the undirected version. In this case, the size of the graph is exponential in the diameter, but only because the degree also grows with d: the size is 2^d, the degree is d (not $2d$, because the generators have order 2), and the diameter is also d.

We will see that, with a fixed small number d of generators, one can obtain nearly optimal results for undirected Abelian Cayley graphs by using a twisted toroidal mesh; the twist allows one to multiply the number of nodes in the ordinary toroidal mesh by 2^{d-1} without increasing the diameter. For two dimensions, we can get exactly optimal results by slightly adjusting this graph; for higher dimensions, the optimal size is still open. For the directed case, we can again get optimal results in two dimensions, but the higher-dimensional case is again unsolved (and rather strange even in three dimensions).

To get these results, we argue as in Proposition 1.1 but focus our argument specifically on the case of Abelian groups. Hence, instead of the free group on d generators, we can use the free Abelian group on d generators, which is simply

$\mathbf{Z}^d$ with the canonical generators $\mathbf{e}_i$, $1 \le i \le d$. For any Abelian group G generated by $g_1, \dots, g_d$, there is a unique homomorphism from $\mathbf{Z}^d$ onto G which sends $\mathbf{e}_i$ to g_i for all i. Let N be the kernel of this homomorphism; then G is isomorphic to $\mathbf{Z}^d/N$, and the Cayley graph of G with the given generators is isomorphic to the Cayley graph of $\mathbf{Z}^d/N$ with the canonical generators for $\mathbf{Z}^d$.

Given a diameter limit k, let S_k be the set of elements of $\mathbf{Z}^d$ which can be expressed as a word of length at most k in the generators $\mathbf{e}_i$, which are allowed to occur positively or negatively. (The dimension d will be clear from the context.) Then S_k can also be described as the set of points in $\mathbf{Z}^d$ at a distance of at most k from the origin under the l^1 (Manhattan) metric:

$$S_k = \{(x_1, \dots, x_d) \in \mathbf{Z}^d : |x_1| + \cdots + |x_d| \le k\}.$$

Let S'_k be the subset of S_k consisting of those elements whose coordinates are all nonnegative; these are the elements which can be expressed as words of length at most k in the generators $\mathbf{e}_i$, where only positive occurrences of the generators are allowed. Then S_k looks like a regular dual d-cube (a square for $d = 2$, an octahedron for $d = 3$), while S'_k looks like a right d-simplex (a triangle for $d = 2$, a tetrahedron for $d = 3$).

Now, by the proof of Proposition 1.1 we get the following.

Proposition 2.1. *Let G, N, and $g_1, \dots, g_d$ be as above. Then the undirected Cayley graph for G and $g_1, \dots, g_d$ has diameter at most k if and only if $S_k + N = \mathbf{Z}^d$, and the directed Cayley graph for G and $g_1, \dots, g_d$ has diameter at most k if and only if $S'_k + N = \mathbf{Z}^d$.*

So $|S_k|$ and $|S'_k|$ give upper bounds for the undirected and directed versions of this case of the degree-diameter problem. It is not hard to show that

$$|S'_k| = \binom{k+d}{d},$$

so $|S'_k| = k^d/d! + O(k^{d-1})$ for fixed d. For $|S_k|$, we easily get the asymptotic form $|S_k| = k^d 2^d/d! + O(k^{d-1})$ for fixed d, but exact formulas are harder; Stanton and Cowan [22] give several, such as

$$|S_k| = \sum_{i=0}^{d} 2^i \binom{d}{i} \binom{k}{i}.$$

In particular, when d is 1, 2, or 3, the formula for $|S_k|$ is $2k + 1$, $2k^2 + 2k + 1$, or $(4k^3 + 6k^2 + 8k + 3)/3$, respectively.

3. Lattice coverings and tilings. Proposition 2.1 tells us that, to find an optimal undirected (directed) Cayley graph of diameter k on d generators, we should look for a subgroup N of $\mathbf{Z}^d$ such that $S_k + N$ ($S'_k + N$) is all of $\mathbf{Z}^d$ and the index $|\mathbf{Z}^d : N|$ is as large as possible; the largest index we can hope for is $|S_k|$ ($|S'_k|$). But the structure of subgroups of $\mathbf{Z}^d$ of finite index (which are all normal, of course) is well known; they are precisely the d-dimensional lattices in $\mathbf{Z}^d$. Because of this, in the rest of the paper we will use the letter L instead of N for such subgroups of $\mathbf{Z}^d$.

A d-dimensional lattice L in $\mathbf{Z}^d$ is specified by d linearly independent vectors $\mathbf{v}_1, \dots, \mathbf{v}_d$ in $\mathbf{Z}^d$; L is the set of all integral linear combinations of these vectors. We have $|\mathbf{Z}^d : L| = |\det M|$, where M is the $d \times d$ matrix whose ith row is $\mathbf{v}_i$ for $i = 1, \dots, d$.

Note that any bounded set contains only finitely many members of L. It follows that, if S is a bounded subset of $\mathbf{Z}^d$ and $\mathbf{x}$ is a point in $\mathbf{Z}^d$, then there are only finitely many $\mathbf{v} \in L$ such that $\mathbf{x} \in S + \mathbf{v}$.

Also note that L, or indeed the entire group $\mathbf{Z}^d$, has a linear ordering $\prec$ which is compatible with addition: $\mathbf{x} \prec \mathbf{y}$ implies $\mathbf{x} + \mathbf{z} \prec \mathbf{y} + \mathbf{z}$. To define $\prec$, first choose a direction (a nonzero vector $\mathbf{v}$ in $\mathbf{R}^d$), and put $\mathbf{x} \prec \mathbf{y}$ if $\mathbf{y}$ is farther in this direction than $\mathbf{x}$ is ($\mathbf{x} \cdot \mathbf{v} < \mathbf{y} \cdot \mathbf{v}$). If two vectors are at the same distance in this direction, then compare them in a second direction; repeat until all ties are broken. One example of this is lexicographic order: compare according to the first coordinates, then according to the second coordinates if the first coordinates are equal, and so on. Or, in this discrete case, one can choose the initial direction so that it distinguishes all points and no tie-breaking is necessary; for instance, if $\mathbf{v} = (1, \pi, \pi^2, \dots, \pi^{d-1})$, then we never have $\mathbf{x} \cdot \mathbf{v} = \mathbf{y} \cdot \mathbf{v}$ for distinct $\mathbf{x}, \mathbf{y}$ in $\mathbf{Z}^d$, so we can just define $\mathbf{x} \prec \mathbf{y}$ to mean $\mathbf{x} \cdot \mathbf{v} < \mathbf{y} \cdot \mathbf{v}$.

A *lattice covering* of $\mathbf{Z}^d$ by a set $S \subseteq \mathbf{Z}^d$ is a collection of translates of S by members of a lattice L (i.e., $\{S + \mathbf{v} : \mathbf{v} \in L\}$) which covers $\mathbf{Z}^d$. If the translates are disjoint, so that each point of $\mathbf{Z}^d$ is covered exactly once, then we have a *lattice tiling* of $\mathbf{Z}^d$ by S.

If we have a lattice covering as above, then $|S| \geq |\mathbf{Z}^d : L|$; if it is a tiling, then $|S| = |\mathbf{Z}^d : L|$. So we can measure the extent to which a covering is "almost" a tiling by one of two numbers: the *density* of the covering, which is $|\mathbf{Z}^d : L| / |S| \geq 1$ (this is the average number of sets in the covering to which a random point of $\mathbf{Z}^d$ belongs), or the *efficiency* of the covering, which is $|S| / |\mathbf{Z}^d : L| \leq 1$. So Proposition 2.1 tells us that, in order to get the best possible Abelian Cayley graph on d generators with diameter k, we must find a lattice covering of $\mathbf{Z}^d$ by S_k or S'_k whose density (efficiency) is as small (large) as possible. We now give one more reformulation of the question.

Lemma 3.1. *Suppose we have a lattice covering of $\mathbf{Z}^d$ using a bounded set S and the lattice L. Then there is a set $T \subseteq S$ such that the translates of T by L form a lattice tiling of $\mathbf{Z}^d$.*

Proof. Let $\prec$ be a linear order of L compatible with addition. Now let T be the set of all points in S which are not in any of the sets $S + \mathbf{v}$ for $\mathbf{v} \in L$, $\mathbf{v} \succ \mathbf{0}$. We will see that every point $\mathbf{x}$ is in exactly one of the sets $T + \mathbf{v}$ for $\mathbf{v} \in L$.

Fix $\mathbf{x}$. As noted before, $\mathbf{x}$ is in only finitely many of the translates $S + \mathbf{v}$, so let $\mathbf{w}$ be the $\prec$-greatest member of L such that $\mathbf{x} \in S + \mathbf{w}$. Let $\mathbf{y} = \mathbf{x} - \mathbf{w} \in S$. Then $\mathbf{y}$ cannot be in $S + \mathbf{v}$ for $\mathbf{v} \succ \mathbf{0}$ in L, because, if it were, we would have $\mathbf{x} = \mathbf{y} + \mathbf{w} \in S + \mathbf{v} + \mathbf{w}$ and $\mathbf{v} + \mathbf{w} \succ \mathbf{w}$, contradicting the maximality of $\mathbf{w}$. So $\mathbf{y} \in T$ and $\mathbf{x} \in T + \mathbf{w}$.

Now suppose $\mathbf{x} = \mathbf{y} + \mathbf{w} = \mathbf{y}' + \mathbf{w}'$, where $\mathbf{w}$ and $\mathbf{w}'$ are distinct members of L and $\mathbf{y}$ and $\mathbf{y}'$ are in T (and hence in S). Then $\mathbf{v} = \mathbf{y} - \mathbf{y}' = \mathbf{w}' - \mathbf{w}$ is a nonzero member of L, so either $\mathbf{v} \succ \mathbf{0}$ or $\mathbf{v} \prec \mathbf{0}$. In the former case, $\mathbf{y} = \mathbf{y}' + \mathbf{v} \in S + \mathbf{v}$ contradicts $\mathbf{y} \in T$; in the latter case, $\mathbf{y}' = \mathbf{y} - \mathbf{v} \in S + (-\mathbf{v})$ contradicts $\mathbf{y}' \in T$. □

Note that such a set T must be of cardinality $|\mathbf{Z}^d : L|$, which is the size of the Cayley graph of $\mathbf{Z}^d/L$. So the size of the largest undirected (directed) Abelian Cayley graph on d generators with diameter k is equal to the size of the largest subset T of S_k (S'_k) such that there is a lattice tiling of $\mathbf{Z}^d$ using T.

4. Approximation by lattice coverings of real space. The study of lattice coverings and lattice tilings is more familiar for $\mathbf{R}^d$ than for $\mathbf{Z}^d$. We will show that real coverings can be approximated to some extent by integer coverings, and vice versa, so that known results from the real context can be transferred to the integer lattices we are interested in.

The definitions of lattice, lattice covering, and lattice tiling are the same in $\mathbf{R}^d$ as in $\mathbf{Z}^d$, except that we allow boundaries to be shared in the definition of a tiling. This lets us work throughout with closed sets (usually polyhedra with their interiors) instead of having to keep some of the boundary points and discard others. Most of the results above for integer lattices go through verbatim for real lattices, including Lemma 3.1 and its proof (although in practice we will probably use the closure of T rather than T itself). The main difference is that the absolute determinant $|\det M|$ of the matrix formed from the generators of a lattice L is not $|\mathbf{R}^d : L|$ (which is infinite). Instead, this determinant is the measure of the parallelepiped determined by the generating vectors; this parallelepiped gives a lattice tiling of $\mathbf{R}^d$ using the lattice L. It follows easily that any other set S which gives a lattice tiling using the lattice L must have measure $|\det M|$ (barring pathological cases of nonmeasurable sets or positive-measure boundaries). Such a set S is called a *fundamental region* for the lattice L. One can now define the density or efficiency of a covering by dividing the measure of the covering set by this determinant or vice versa.

One can transform a lattice covering of $\mathbf{Z}^d$ using S into a lattice covering of $\mathbf{R}^d$ by replacing each point of S with a unit cube (i.e., replace S with $S+U$, where U is a fixed unit d-cube with edges parallel to the coordinate axes); the two coverings will have the same density. However, transforming results in the other direction is harder because the real results usually involve actual triangles, octahedra, etc., rather than polycube approximations, and the lattices used often will not be integer lattices. We will now present results that allow us to get around these difficulties.

In $\mathbf{R}^d$, let $\bar{S}_k$ be the closed l^1-ball of radius k at the origin:

$$\bar{S}_k = \{(x_1,\dots,x_d) : |x_1| + \cdots + |x_d| \le k\}.$$

Let $\bar{S}'_k$ be the set of nonnegative points in $\bar{S}_k$:

$$\bar{S}'_k = \{(x_1,\dots,x_d) : x_1,\dots,x_d \ge 0,\ x_1 + \cdots + x_d \le k\}.$$

Let L be any lattice in $\mathbf{R}^d$.

Proposition 4.1.

(a) *If $S_k + L$ covers $\mathbf{Z}^d$, then $\bar{S}_{k+d/2} + L$ covers $\mathbf{R}^d$.*

(b) *If $S'_k + L$ covers $\mathbf{Z}^d$, then $\bar{S}'_{k+d} + L$ covers $\mathbf{R}^d$.*

Proof. (a) By the triangle inequality for l^1 distance, we have $\bar{S}_k + \bar{S}_{d/2} \subseteq \bar{S}_{k+d/2}$, so $S_k + L + \bar{S}_{d/2} \subseteq \bar{S}_{k+d/2} + L$; therefore, it suffices to show that $\mathbf{Z}^d + \bar{S}_{d/2} = \mathbf{R}^d$. For any $\mathbf{x} \in \mathbf{R}^d$, let $\mathbf{y}$ be the element of $\mathbf{Z}^d$ nearest to $\mathbf{x}$ (i.e., round each coordinate of $\mathbf{x}$ to the nearest integer); then $\mathbf{x} - \mathbf{y}$ is in $\bar{S}_{d/2}$, so $\mathbf{x}$ is in $\mathbf{Z}^d + \bar{S}_{d/2}$.

(b) Similarly, this follows from the fact that $\mathbf{Z}^d + \bar{S}'_d = \mathbf{R}^d$, which is proved in the same way as above (round each coordinate of $\mathbf{x}$ downward instead of to the nearest integer). □

One can argue in the same way within $\mathbf{Z}^d$ to get the following.

Proposition 4.2. *If L is a lattice in $\mathbf{Z}^d$ and m is a positive integer, then the following hold:*

(a) *If $S_k + L$ covers $\mathbf{Z}^d$, then $S_{mk+\lfloor m/2 \rfloor d} + mL$ covers $\mathbf{Z}^d$.*

(b) *If $S'_k + L$ covers $\mathbf{Z}^d$, then $S'_{mk+(m-1)d} + mL$ covers $\mathbf{Z}^d$.*

Proof. For (a), clearly $S_{mk} + mL$ covers $m\mathbf{Z}^d$, so, as in the preceding proposition, it suffices to note that $m\mathbf{Z}^d + S_{\lfloor m/2 \rfloor d}$ covers $\mathbf{Z}^d$ because we can simply round any member of $\mathbf{Z}^d$ to the nearest member of $m\mathbf{Z}^d$. Similarly, (b) holds because $m\mathbf{Z}^d + S'_{(m-1)d}$ covers $\mathbf{Z}^d$. □

Of course, one can get similar results for sets other than S_k and S'_k.

Using Proposition 4.1, it is easy to move from a covering of $\mathbf{Z}^d$ using an integer lattice to a covering of $\mathbf{R}^d$ using a real lattice (and using the real shape $\bar{S}_k$ or $\bar{S}'_k$); if k is large relative to d, then the two coverings have about the same efficiency. We will now show that one can move in the other direction as well.

First, we give a useful criterion for deciding whether one has a lattice covering of $\mathbf{R}^d$.

Proposition 4.3. *Suppose S is a nonempty subset of $\mathbf{R}^d$ and L is a lattice in $\mathbf{R}^d$. If there is an $\varepsilon > 0$ such that $S + L$ covers all points within distance ε of S, then $S + L$ covers $\mathbf{R}^d$.*

Proof. Clearly, there is some point $\mathbf{x}_0$ in $S + L$. We will show that, if $\mathbf{x} \in S + L$ and the distance $\delta(\mathbf{x}, \mathbf{y})$ is less than ε, then $\mathbf{y} \in S + L$. Applying this once shows that all points within distance ε of $\mathbf{x}_0$ are in $S + L$; applying it again shows that all points within distance 2ε of $\mathbf{x}_0$ are in $S + L$; since this can be repeated ad infinitum, we find that all points of $\mathbf{R}^d$ are in $S + L$.

Let $\mathbf{x}$ and $\mathbf{y}$ be as above. Find $\mathbf{v} \in L$ such that $\mathbf{x} \in S + \mathbf{v}$. Then $\mathbf{y} - \mathbf{v}$ is within distance ε of $\mathbf{x} - \mathbf{v} \in S$, so there is $\mathbf{v}' \in L$ such that $\mathbf{y} - \mathbf{v} \in S + \mathbf{v}'$. Hence, $\mathbf{y} \in S + \mathbf{v}' + \mathbf{v}$, so $\mathbf{y} \in S + L$, as desired. □

Using this, we can now show that, if one has a lattice covering using a bounded subset of $\mathbf{R}^d$, then one can perturb the lattice slightly and still get a lattice covering using a slightly larger subset of $\mathbf{R}^d$.

Proposition 4.4. *Let S be a bounded subset of $\mathbf{R}^d$, and let L be a lattice in $\mathbf{R}^d$ such that $S + L = \mathbf{R}^d$; let $\mathbf{v}_1, \ldots, \mathbf{v}_d$ be a list of generators for L. Then there are positive numbers η and ρ such that, for all $r \in (0, 1)$, if the distance $\delta(\mathbf{v}_i, \mathbf{v}'_i)$ is less than $r\eta$ for all $i \le d$, then $S^+ + L' = \mathbf{R}^d$, where L' is the lattice generated by $\mathbf{v}'_1, \ldots, \mathbf{v}'_d$ and S^+ is the set of points within distance $r\rho$ of S.*

Proof. The number of members of L within any bounded part of $\mathbf{R}^d$ is finite, so it only takes finitely many translates of S by members of L to cover any bounded part of $\mathbf{R}^d$. In particular, there is a number $M > 0$ such that the sets $S + a_1\mathbf{v}_1 + \cdots + a_d\mathbf{v}_d$ for $(a_1, \ldots, a_d) \in \mathbf{Z}^d$ with $|a_1| + \cdots + |a_d| \le M$ cover all points within distance ρ of S, for some $\rho > 0$. Let $\eta = \rho/M$.

Let r be any positive number less than 1; we must see that, if S^+ and L' are defined as above, then $S^+ + L' = \mathbf{R}^d$. By the preceding proposition, it will suffice to show that $S^+ + L'$ covers all points within distance $(1 - r)\rho$ of S^+. Suppose $\mathbf{y}$ is within distance $(1 - r)\rho$ of S^+; then $\mathbf{y}$ is within distance $(1 - r)\rho + r\rho = \rho$ of S, so there exist integers $a_1, \ldots, a_d$ with $|a_1| + \cdots + |a_d| \le M$ and a point $\mathbf{x} \in S$ such that $\mathbf{y} = \mathbf{x} + a_1\mathbf{v}_1 + \cdots + a_d\mathbf{v}_d$. Let $\mathbf{x}' = \mathbf{x} + \sum_{i=1}^d a_i(\mathbf{v}_i - \mathbf{v}'_i)$; then

$$\delta(\mathbf{x}, \mathbf{x}') \le \sum_{i=1}^d |a_i|\delta(\mathbf{v}_i, \mathbf{v}'_i) < \sum_{i=1}^d |a_i| r\eta \le Mr\eta = r\rho,$$

so $\mathbf{x}' \in S^+$. Since $\mathbf{y} = \mathbf{x}' + a_1\mathbf{v}'_1 + \cdots + a_d\mathbf{v}'_d$, we have $\mathbf{y} \in S^+ + L'$, as desired. □

Note that, in the above propositions, "distance" need not be Euclidean distance; it can be any metric arising from a norm on $\mathbf{R}^d$. For our present purposes, it will be most convenient (but not essential) to use l^∞-distance: $\delta(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i|$.

Theorem 4.5. *Suppose one has a lattice L in $\mathbf{R}^d$ such that $\bar{S}_k + L$ covers $\mathbf{R}^d$; let $\mathbf{v}_1, \ldots, \mathbf{v}_d$ be generators for L. Then there is a constant c such that, for all sufficiently large real numbers t, if $\mathbf{w}_i$ is obtained from $t\mathbf{v}_i$ by rounding all coordinates to the nearest integer, and $\bar{L}$ is the lattice generated by $\mathbf{w}_1, \ldots, \mathbf{w}_d$, then $\bar{S}_{tk+c} + \bar{L}$ covers $\mathbf{R}^d$. The same statement holds for $\bar{S}'_k$ and $\bar{S}'_{tk+c}$ instead of $\bar{S}_k$ and $\bar{S}_{tk+c}$.*

Proof. For the $\bar{S}_k$ case, let $S = \bar{S}_k$ and find η and ρ as in the preceding proposition, letting the distance δ be the l^∞ metric. Let c be any fixed number greater than $d\rho/2\eta$. Then, for any $t > c/d\rho$, if we let $r = c/d\rho t$, then $r < 1$ and $1/2t < r\eta$. If we define $\mathbf{w}_i$ and $\bar{L}$ as above, and let $\mathbf{v}'_i = \mathbf{w}_i/t$ and $L' = \bar{L}/t$, then $\delta(\mathbf{v}_i, \mathbf{v}'_i) \le 1/2t$ for each i, so we can conclude that $S^+ + L'$ covers $\mathbf{R}^d$, where S^+ is the set of points within distance $r\rho$ of $\bar{S}_k$. It is easy to see that $S^+ \subseteq \bar{S}_{k+dr\rho}$, so $\bar{S}_{k+dr\rho} + L'$ covers $\mathbf{R}^d$. Hence, $\bar{S}_{tk+tdr\rho} + tL'$ covers $\mathbf{R}^d$; but $tdr\rho = c$ and $tL' = \bar{L}$, so we are done.

The proof for $\bar{S}'_k$ is almost the same. Let $S = \bar{S}'_k$ and apply the preceding proposition (using the l^∞ metric again) to get η and ρ. Fix $c > d\rho/\eta$. For any $t > c/2d\rho$, if we let $r = c/2d\rho t$, then $r < 1$ and $1/2t < r\eta$. Now define $\mathbf{w}_i$, $\bar{L}$, $\mathbf{v}'_i$, L', and S^+ as above, and conclude again that $S^+ + L'$ covers $\mathbf{R}^d$. One can easily check that $S^+ \subseteq \bar{S}'_{k+2dr\rho} - r\rho\mathbf{u}$, where $\mathbf{u} = (1, \dots, 1)$. Hence, $\bar{S}'_{k+2dr\rho} - r\rho\mathbf{u} + L'$ covers $\mathbf{R}^d$, so $\bar{S}'_{k+2dr\rho} + L'$ covers $\mathbf{R}^d + r\rho\mathbf{u} = \mathbf{R}^d$. Now multiply by t to see that $\bar{S}'_{tk+c} + \bar{L}$ covers $\mathbf{R}^d$. □

Again it is easy to modify this proof for other sets in place of $\bar{S}_k$ or $\bar{S}'_k$. Also, the proof is quite effective, allowing one to compute specific values of c and t which work for a given lattice L (assuming it is feasible to compute M and ρ).

The covering $\bar{S}_{tk}+tL$ has the same efficiency as the covering $\bar{S}_k+L$; since $\bar{S}_{tk+c}+\bar{L}$ is a relatively slight perturbation of $\bar{S}_{tk} + tL$ when t is large, it has almost the same efficiency. Since $\bar{L}$ is an integer lattice, the fact that $\bar{S}_{tk+c} + \bar{L}$ covers $\mathbf{R}^d$ implies that $S_{\lfloor tk+c \rfloor} + \bar{L}$ covers $\mathbf{Z}^d$; again the efficiency is almost the same if t is large. Therefore, we can construct integer lattice coverings which are as nearly efficient as desired to a given real lattice covering, thus giving asymptotic results for the present case of the degree-diameter problem. The precise result is as follows.

Theorem 4.6. *Let $\varepsilon_{\mathbf{R}}$ be the best possible efficiency for a lattice covering of $\mathbf{R}^d$ by $\bar{S}_1$, and let $\varepsilon_{\mathbf{Z}}(k)$ be the best possible efficiency for a lattice covering of $\mathbf{Z}^d$ by S_k. Then $\varepsilon_{\mathbf{Z}}(k) = \varepsilon_{\mathbf{R}} + O(k^{-1})$. The same applies to $\bar{S}'_1$ and S'_k.*

Proof. Let L be a lattice giving a lattice covering of $\mathbf{R}^d$ by $\bar{S}_1$ with efficiency $\varepsilon_{\mathbf{R}}$. Let $\mathbf{v}_1, \dots, \mathbf{v}_d$ and c be as in Theorem 4.5. Given a large integer k, let $t = k-c$, and let $\bar{L}$ be the integer lattice approximating tL as in Theorem 4.5, generated by $\mathbf{w}_1, \dots, \mathbf{w}_d$. Since $\bar{L}$ is an integer lattice and $\bar{S}_{t+c} + \bar{L} = \mathbf{R}^d$, we have $S_k + \bar{L} = \mathbf{Z}^d$. Let M and $\bar{M}$ be the $d \times d$ matrices whose rows are $\mathbf{v}_i$ and $\mathbf{w}_i$, respectively; then $\det M = (2^d/d!)\varepsilon_{\mathbf{R}}$ and $\det \bar{M} \le |S_k|\varepsilon_{\mathbf{Z}}(k)$, which implies $\det(k^{-1}\bar{M}) \le (2^d/d! + O(k^{-1}))\varepsilon_{\mathbf{Z}}(k)$. But $\bar{M} = tM + O(1) = kM + O(1)$, so $k^{-1}\bar{M} = M + O(k^{-1})$, and thus $\det(k^{-1}\bar{M}) = \det M + O(k^{-1})$; this implies $\varepsilon_{\mathbf{Z}}(k) \ge \varepsilon_{\mathbf{R}} + O(k^{-1})$.

On the other hand, Proposition 4.1(a) states that any lattice that gives a covering of $\mathbf{Z}^d$ by S_k also gives a covering of $\mathbf{R}^d$ by $\bar{S}_{k+d/2}$. If the efficiency of the former is $\varepsilon_{\mathbf{Z}}(k)$, then the efficiency of the latter is $(|S_k|/\operatorname{vol}(\bar{S}_{k+d/2}))\varepsilon_{\mathbf{Z}}(k) = (1+O(k^{-1}))\varepsilon_{\mathbf{Z}}(k)$; hence, $(1 + O(k^{-1}))\varepsilon_{\mathbf{Z}}(k) \le \varepsilon_{\mathbf{R}}$, so $\varepsilon_{\mathbf{Z}}(k) \le \varepsilon_{\mathbf{R}} + O(k^{-1})$.

The same argument works for $\bar{S}'_1$ and S'_k. □

Again the same applies to other shapes as well. Combining this with the known sizes of the sets S_k and S'_k gives the following.

Corollary 4.7. *Let d be a fixed positive integer.*

(a) *If $\varepsilon_{\mathbf{R}}$ is the best possible efficiency for a lattice covering of $\mathbf{R}^d$ by $\bar{S}_1$, then the size of the largest possible undirected Cayley graph of an Abelian group on d generators with diameter at most k is $(2^d\varepsilon_{\mathbf{R}}/d!)k^d + O(k^{d-1})$.*

(b) *If $\varepsilon'_{\mathbf{R}}$ is the best possible efficiency for a lattice covering of $\mathbf{R}^d$ by $\bar{S}'_1$, then the size of the largest possible directed Cayley graph of an Abelian group on d generators with diameter at most k is $(\varepsilon'_{\mathbf{R}}/d!)k^d + O(k^{d-1})$.*

The above assumes that the upper limit $\varepsilon_{\mathbf{R}}$ on the efficiency of a lattice covering of $\mathbf{R}^d$ by $\bar{S}_1$ is actually attained. To see that this is the case, first note that there is certainly a lattice covering with positive efficiency, say ε_0. Now consider the possible sets of generating vectors $\mathbf{v}_1, \dots, \mathbf{v}_d$ for a lattice L giving a covering of efficiency at least ε_0. We may assume that, if $\mathbf{v}_i$ is one of the generators and $\mathbf{w}$ is an integral linear combination of the other generators, then $|\mathbf{v}_i| \le |\mathbf{v}_i + \mathbf{w}|$; otherwise, just replace $\mathbf{v}_i$ with $\mathbf{v}_i + \mathbf{w}$ to get a smaller set of generating vectors for L, and iterate until no more such reductions are possible. It follows that, if, say $\mathbf{v}_1$ is the longest of the vectors $\mathbf{v}_i$, then the angle between $\mathbf{v}_1$ and the hyperplane P spanned by $\mathbf{v}_2, \dots, \mathbf{v}_d$ is bounded below by a positive number ($\pi/3$ for $d = 2$, somewhat less for higher d). But the distance from $\mathbf{v}_1$ to P is at most the diameter of $\bar{S}_1$ (i.e., 2) since otherwise $\bar{S}_1 + L$ would consist of “hyperplanes” of copies of $\bar{S}_1$ with gaps in between. Putting these together, we get a fixed upper bound B on the length $|\mathbf{v}_1|$, and hence on all of the lengths $\mathbf{v}_i$. But we also have a positive lower bound b on the determinant $\det(\mathbf{v}_1, \dots, \mathbf{v}_d)$, namely $\varepsilon_0 \operatorname{vol}(\bar{S}_1)$. One can now show that there is a fixed number M such that, if $m_1, \dots, m_d \in \mathbf{Z}$ and $|m_1| + \cdots + |m_d| > M$, then $|m_1\mathbf{v}_1 + \cdots + m_d\mathbf{v}_d| > 3$. Hence, the finitely many translates $\bar{S}_1 + m_1\mathbf{v}_1 + \cdots + m_d\mathbf{v}_d$ with $|m_1| + \cdots + |m_d| \le M$ will have to cover all points at distance ≤ 2 from the origin. Now, the set of sequences $\mathbf{v}_1, \dots, \mathbf{v}_d$ with all $|\mathbf{v}_i| \le B$ such that the translates $\bar{S}_1 + m_1\mathbf{v}_1 + \cdots + m_d\mathbf{v}_d$ for $|m_1| + \cdots + |m_d| \le M$ cover all points at distance ≤ 2 from $\mathbf{0}$ is a compact set, so there is a sequence $\mathbf{v}_1, \dots, \mathbf{v}_d$ in this set for which $\det(\mathbf{v}_1, \dots, \mathbf{v}_d)$ is maximal; this sequence of vectors generates a lattice covering of $\mathbf{R}^d$ by $\bar{S}_1$ (by Proposition 4.3) with maximal possible efficiency. The same argument works for any compact shape of positive volume, such as $\bar{S}'_1$. (Presumably this argument is well known, but the authors were not able to find a reference for it.)

As we will see in a later section (for the specific case $d = 3$, but by a general argument), the determination of actual values for $\varepsilon_{\mathbf{R}}$ and $\varepsilon'_{\mathbf{R}}$ for a particular dimension d can, in principle, be reduced to the solution of a finite number of nonlinear optimization problems (each of which requires maximizing a degree-d polynomial function over a region which is a convex polytope in $\mathbf{R}^{d^2}$). Unfortunately, this finite number is extremely large, so the actual values are not known for $d > 2$. (For $d = 1$ we trivially have $\varepsilon_{\mathbf{R}} = \varepsilon'_{\mathbf{R}} = 1$. For $d = 2$ we will see below that $\varepsilon_{\mathbf{R}} = 1$ and $\varepsilon'_{\mathbf{R}} = 2/3$.) The computations described in later sections lead us to conjecture that, for $d = 3$, $\varepsilon_{\mathbf{R}} = 8/9$ and $\varepsilon'_{\mathbf{R}} = 63/125$.

5. Undirected Cayley graphs on two generators. We now begin to consider the results that can be obtained for specific values of d. As noted previously, the case $d = 1$ is trivial, so we will start with the case of undirected Abelian Cayley graphs on two generators and a given bound k on the diameter. The results in this subsection are not new.

As noted earlier, the diameter of the ordinary toroidal mesh $\mathbf{Z}_m \times \mathbf{Z}_n$ is $\lfloor m/2 \rfloor + \lfloor n/2 \rfloor$. Hence, the largest such mesh with diameter $\le k$ is the one in which m is k rounded up to the nearest odd integer and n is $k + 1$ rounded up to the nearest odd integer. This corresponds to the lattice covering of $\mathbf{Z}^2$ by S_k using the lattice $L = m\mathbf{Z} \times n\mathbf{Z}$; the efficiency of this covering is $mn/(2k^2 + 2k + 1)$, which tends to $1/2$ as k becomes large. So one can hope for better results.

To get such results, consider the real rotated square $\bar{S}_k$. There is, obviously, a lattice tiling using this square; it is just a rotated orthogonal grid with spacing $\sqrt{2}k$ between lines. The lattice L_1 for this tiling is generated by the vectors (k, k) and

$(-k,k)$; as expected, the corresponding determinant is $2k^2$, which is equal to the area of $\bar{S}_k$.

It now follows from the approximation results that we can get lattice coverings of $\mathbf{Z}^2$ using S_k with efficiencies that approach 1 for large k. However, we do not need the general results here; since L_1 is already an integer lattice, we can simply note that $\bar{S}_k + L_1 = \mathbf{R}^2$ implies $S_k + L_1 = \mathbf{Z}^d$. So this gives a lattice covering using S_k whose index is $|\mathbf{Z}^2 : L_1| = 2k^2$, which is better than that from the best toroidal mesh for all $k \geq 3$; for large k, the efficiency approaches 1.

The corresponding Cayley graph $\mathbf{Z}^2/L_1$ turns out to be quite simple to describe. The $2k \times k$ rectangle $\{1, \ldots, 2k\} \times \{1, \ldots, k\}$ contains exactly one point from each coset of L_1, so it can serve as a set of vertices for the graph. Adjacent points in the rectangle (horizontally and vertically) are connected as in the usual mesh. Horizontally, one has the usual toroidal connections at the ends: $(1,j)$ is connected to $(2k,j)$. But vertically, there is an offset of k: $(i,1)$ is connected to $(i+k,k)$ if $i \leq k$, or to $(i-k,k)$ if $i > k$. This is just like a $2k \times k$ toroidal mesh, except that the torus is twisted halfway around before the long edges are glued together. This twist allows one to double the number of vertices in a $k \times k$ toroidal mesh while increasing the diameter by at most 1 (there is no increase if k is even).

A number of the useful properties of ordinary toroidal meshes apply with very little change to twisted toroidal meshes. For instance, since the new mesh is still just a rectangular mesh with extra connections at the boundary, it is easy to map a simple rectangular grid into the mesh by simply ignoring the boundary connections.

Another nice property of toroidal meshes is that it is easy to find a shortest route from one node to another: just check each coordinate separately to find which of the two possible directions gives a shorter path and put the results together. Finding optimal routes is only slightly more complicated for the twisted toroidal mesh. To see this, consider the given $2k \times k$ rectangle as half of a $2k \times 2k$ rectangle: each node (i,j) in the first half has a copy $(i \pm k, j \pm k)$ in the other half. This larger rectangle is then copied periodically without further twists to cover $\mathbf{Z}^2$; in other words, the $2k \times k$ twisted toroidal mesh is just a $2k \times 2k$ toroidal mesh where (i,j) and $(i \pm k, j \pm k)$ are identified as a single node. Therefore, to find an optimal route from (i,j) to (i',j') in the twisted mesh, apply the ordinary $2k \times 2k$ toroidal mesh routing algorithm to find optimal routes from (i,j) to (i',j') and to $(i' \pm k, j' \pm k)$, and choose the shorter of the two.

In the real case, the lattice L_1 gave a perfect tiling of $\mathbf{R}^2$ using $\bar{S}_k$, since boundary overlap did not count; but in the integer case, the boundary overlap reduces the efficiency slightly from 1 to $2k^2/(2k^2+2k+1)$. It turns out that if one uses a slightly modified lattice, namely the lattice L_2 with generating vectors $(k,k+1)$ and $(-k-1,k)$, then one gets a covering of $\mathbf{Z}^2$ by copies of S_k with efficiency 1 (i.e., a tiling). See Figure 5.1. We therefore get the following.

Theorem 5.1 (multiple authors; see below). *The largest possible size for the undirected Cayley graph of an Abelian group on two generators with diameter k is $2k^2+2k+1$.*

This result has appeared in various forms in a number of places (usually stated so as to apply only to cyclic Cayley graphs, but since the optimal Abelian Cayley graphs turn out to be cyclic, the results are basically equivalent). See, for instance, Boesch and Wang [4] or Yebra el al. [25]; the Bermond–Comellas–Hsu survey [3] also has many additional references.

The tiling in Figure 5.1 appears in Yebra et al. [25], among other places; it

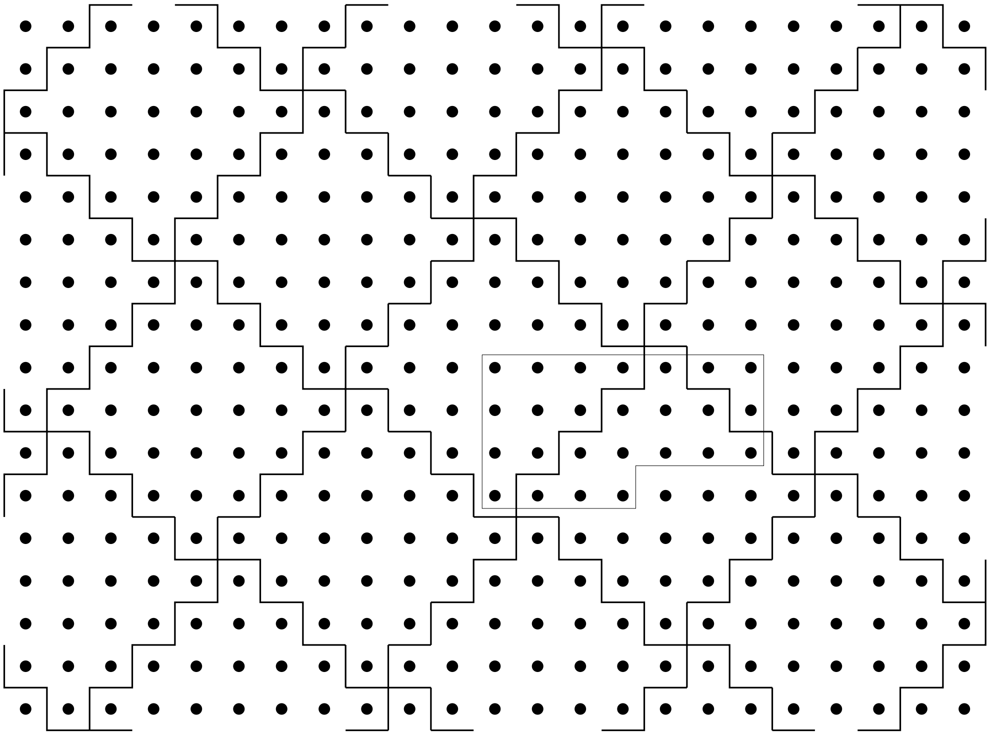


FIG. 5.1. *Lattice tiling of* $\mathbf{Z}^2$ *using* S_k *(shown for* $k = 3$*).*

even appears in Native American artwork of the southwestern United States and may date back to the ancient Aztecs, who used the stepped diamond shape in temple ornamentation. (This shape is now commonly known as the Aztec diamond, a term coined by J. Propp.) However, it is unlikely that the Aztecs were motivated by the desire to construct efficient parallel computation networks.

It is easy to see that the lattice tiling of $\mathbf{Z}^2$ by S_k, or of $\mathbf{R}^2$ by the Aztec diamond, is unique except for a possible reflection in the line $x = y$; this just corresponds to interchanging the two generators for the Cayley graph. Therefore, the Cayley graph attaining the bound in Theorem 5.1 is unique up to isomorphism.

Since the point $(2k + 1, 1)$ is in L_2, we have $\mathbf{e}_2 + L_2 = (-2k - 1)(\mathbf{e}_1 + L_2)$ in $\mathbf{Z}^2/L_2$. Hence, $\mathbf{Z}^2/L_2$ is a cyclic group generated by $\mathbf{e}_1 + L_2$ alone. It is isomorphic (not only as a group but as a Cayley graph) to $\mathbf{Z}_{2k^2+2k+1}$ with the generating set $\{1, 2k^2\}$. One may choose to replace the second generator with its inverse, making the generating set $\{1, 2k + 1\}$; other generating sets can be used as well.

For layout purposes, one may just arrange the nodes in the form of the diamond S_k and connect the boundary nodes as specified by L_2, but it is probably more convenient to use the almost-rectangular shape outlined in Figure 5.1 (a $(2k + 1) \times k$ rectangle with an extra partial row of length $k + 1$). The boundary connections are similar to those for the twisted toroidal mesh given earlier, but now there is also a slight twist when connecting the short sides: there is a drop of one row when wrapping around from right to left. This layout shows that one can embed a rectangular grid into this graph so as to use almost all of the nodes.

6. Directed Cayley graphs on two generators. We now describe the largest possible directed Cayley graph of an Abelian group on two generators with diameter bounded by k. As in the preceding subsection, the two-generator results here are already known.

The best toroidal mesh in this case is $\mathbf{Z}_m \times \mathbf{Z}_{m'}$, where $m = \lfloor k/2 \rfloor + 1$ and $n = \lceil k/2 \rceil + 1$; this gives size $mm' = \lfloor (k+2)^2/4 \rfloor$, which is about $1/2$ of $|S'_k|$, so one can hope to do better.

However, one is not going to get perfect efficiency in this case. One can easily tile the plane with triangles such as $\bar{S}'_k$ if one is allowed to rotate them, but this is not possible using only a lattice of translated copies of a triangle. The exact minimum density for a lattice covering of the plane by triangles was computed by Fáry in 1950; we will give a different proof of his result here and then give the analogue for $\mathbf{Z}^2$.

THEOREM 6.1 (Fáry [13]). *The minimum density for a lattice covering of* $\mathbf{R}^2$ *by triangles is* $3/2$. *Equivalently, the maximum efficiency is* $2/3$.

Proof. Since the density and efficiency of a lattice covering are invariant under affine transformations, it does not matter which triangle we work with, so, for slight convenience, let us work with the isosceles right triangle $\bar{S}'_1$.

One can attain the efficiency $2/3$ by using the lattice with generating vectors $(1/3, 1/3)$ and $(2/3, -1/3)$. This corresponds to a tiling of the plane using an L-tromino that takes up $2/3$ of $\bar{S}'_1$, as shown in Figure 6.1. Or one can cut off all three corners of the triangle to get a hexagon that tiles the plane.

Now suppose that we have a lattice covering of $\mathbf{R}^2$ using $\bar{S}'_1$ and the lattice L; we must show that the efficiency of the covering is at most $2/3$. Let $\prec$ be a linear ordering of L compatible with addition which is defined by primarily ordering points (x, y) according to the sum $x + y$ (so points that are farther out in the direction $(1, 1)$ come later in the ordering) and breaking ties (if any) by distance in some other direction.

Let $\overline{AB}$ be the hypotenuse of $\bar{S}'_1$. Only finitely many of the L-translates of $\bar{S}'_1$ lie near $\bar{S}'_1$; of these, those of form $\bar{S}'_1 + \mathbf{v}$ for $\mathbf{v} \succ \mathbf{0}$ must cover the points which are near $\overline{AB}$ on the side facing away from $\bar{S}'_1$. Since the union of finitely many translates of $\bar{S}'_1$ is closed, $\overline{AB}$ itself is covered by finitely many translates $\bar{S}'_1 + \mathbf{v}$ with $\mathbf{v} \succ \mathbf{0}$. Find such a covering of $\overline{AB}$ with as few translates as possible, say $\bar{S}'_1 + \mathbf{v}_1, \dots, \bar{S}'_1 + \mathbf{v}_m$, where $\mathbf{v}_i \succ \mathbf{0}$.

Note that, since $\bar{S}'_1 + \mathbf{v}_i$ must intersect $\overline{AB}$, it contains one of the endpoints A and B if and only if the coordinates of $\mathbf{v}_i$ are not both positive. We may assume that at most one of the vectors $\mathbf{v}_i$ has both coordinates positive. If there are two such vectors, let them be $\mathbf{v}_i$ and $\mathbf{v}_j$, where $\mathbf{v}_i \prec \mathbf{v}_j$. Then $\mathbf{v}_j - \mathbf{v}_i \succ \mathbf{0}$; since $\mathbf{v}_i$ has both

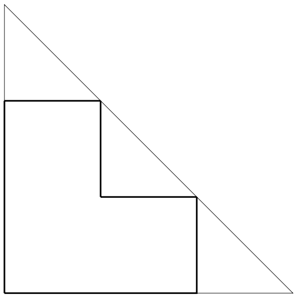


FIG. 6.1. *A subset of the triangle* $\bar{S}'_1$ *which tiles the plane.*

coordinates positive, we have

$$(\bar{S}_1' + \mathbf{v}_j) \cap \overline{AB} \subseteq (\bar{S}_1' + \mathbf{v}_j - \mathbf{v}_i) \cap \overline{AB}.$$

Furthermore, $\mathbf{v}_j - \mathbf{v}_i$ cannot have both coordinates positive because, if it did, we would have

$$(\bar{S}_1' + \mathbf{v}_j) \cap \overline{AB} \subseteq (\bar{S}_1' + \mathbf{v}_i) \cap \overline{AB},$$

so $\bar{S}_1' + \mathbf{v}_j$ would not have been needed in the covering of $\overline{AB}$, contradicting the minimality of m. So we can replace $\bar{S}_1' + \mathbf{v}_j$ with $\bar{S}_1' + \mathbf{v}_j - \mathbf{v}_i$ to get another covering of $\overline{AB}$ using fewer vectors with both coordinates positive. Repeat this until only one such vector is left.

Since each translate $\bar{S}_1' + \mathbf{v}_i$ is convex, its intersection with $\overline{AB}$ is a segment or a point. Therefore, at most one of these translates can contain A, since otherwise one of the intersections $(\bar{S}_1' + \mathbf{v}_i) \cap \overline{AB}$ would include another such intersection, making the latter translate superfluous and contradicting the minimality of m. Similarly, at most one of the translates $\bar{S}_1' + \mathbf{v}_i$ contains B. Putting these facts together, we conclude that we need at most three of the translates $\bar{S}_1' + \mathbf{v}$ with $\mathbf{v} \succ \mathbf{0}$ to cover the segment $\overline{AB}$.

In other words, there are points P and Q on $\overline{AB}$ such that each of the three segments $\overline{AP}$, $\overline{PQ}$, and $\overline{QB}$ is covered by one of the translates $\bar{S}_1' + \mathbf{v}$ with $\mathbf{v} \succ \mathbf{0}$. Let l_1, l_2, l_3 be the lengths of these three segments; then $l_1 + l_2 + l_3 = \sqrt{2}$. Note that, if $\bar{S}_1' + \mathbf{v}$ covers $\overline{AP}$, then $\bar{S}_1' + \mathbf{v}$ covers the entire isosceles right triangle below $\overline{AP}$ whose hypotenuse is $\overline{AP}$; the area of this triangle is $l_1^2/4$. Similar statements hold for $\overline{PQ}$ and $\overline{QB}$. So we have three disjoint triangles included in $\bar{S}_1'$ which are covered by translates $\bar{S}_1' + \mathbf{v}$ with $\mathbf{v} \succ \mathbf{0}$, and the total area of these triangles is $(l_1^2 + l_2^2 + l_3^2)/4$.

By the proof of (the real version of) Lemma 3.1, if we let T be the part of $\bar{S}_1'$ which is not covered by any translate $\bar{S}_1' + \mathbf{v}$ with $\mathbf{v} \succ \mathbf{0}$, then T gives a lattice tiling of $\mathbf{R}^2$ using L, so the efficiency of the covering using $\bar{S}_1'$ and L is $\operatorname{Area}(T)/\operatorname{Area}(\bar{S}_1')$. We have $\operatorname{Area}(\bar{S}_1') = 1/2$ and $\operatorname{Area}(T) \leq 1/2 - (l_1^2 + l_2^2 + l_3^2)/4$. A standard minimization shows that, if $l_1 + l_2 + l_3 = \sqrt{2}$, then $l_1^2 + l_2^2 + l_3^2 \geq 2/3$ (with equality only when $l_1 = l_2 = l_3 = \sqrt{2}/3$). Therefore, $\operatorname{Area}(T) \leq 1/3$, so the efficiency of the covering by $\bar{S}_1'$ and L is at most $2/3$. ∎

It now follows from Theorem 4.6 that the largest possible index $|\mathbf{Z}^2 : L|$ for an integer lattice L giving a lattice covering of $\mathbf{Z}^2$ by S_k' is approximately $(2/3)|S_k'|$, or about $k^2/3$, for large k. However, we can actually get an exact answer rather than an approximation.

THEOREM 6.2 (due mainly to Wong and Coppersmith [24]). *The largest possible index $|\mathbf{Z}^2 : L|$ for a lattice L giving a lattice covering of $\mathbf{Z}^2$ by S_k' is $\lfloor (k+2)^2/3 \rfloor$.*

Proof. We will give a discrete form of the proof of Theorem 6.1. Let a be $(k+2)/3$ rounded to the nearest integer, and let $b = k + 2 - 2a$ (so b is also about $(k+2)/3$). Let T_k be the set of (i, j) in $\mathbf{Z}^2$ such that $i, j \geq 0$, $\min(i, j) < a$, and $\max(i, j) < a + b$. Then $T_k \subseteq S_k'$, since any (i, j) in T_k satisfies $i + j \leq a - 1 + a + b - 1 = k$. The set T_k looks like the L-tromino from Figure 6.1, and it tiles $\mathbf{Z}^2$ using the lattice with generating vectors (a, a) and $(a + b, -b)$. Therefore, this lattice gives a covering of $\mathbf{Z}^2$ using S_k', and its index is $a(a + 2b)$, which works out to be $\lfloor (k+2)^2/3 \rfloor$.

Now, suppose we have a lattice covering of $\mathbf{Z}^2$ using S_k' and a lattice L; we must show that $|\mathbf{Z}^2 : L| \leq \lfloor (k+2)^2/3 \rfloor$. Define the linear order $\prec$ of L as before. Let A and B be the points $(0, k+1)$ and $(k+1, 0)$; then the segment $\overline{AB}$ contains $k+2$ integral points, which must be covered by translates $S_k' + \mathbf{v}$ where $\mathbf{v} \in L$ and $\mathbf{v} \succ \mathbf{0}$.

Let $\mathbf{v}_1, \dots, \mathbf{v}_m$ be a list of as few vectors as possible in L such that $\mathbf{v}_i \succ \mathbf{0}$ and the translates $S'_{k+1} + \mathbf{v}_i$ of S'_{k+1} cover all of the integral points on $\overline{AB}$. Then the same argument as for Theorem 6.1 shows that m is at most 3. Hence, $\overline{AB}$ can be broken up into three segments $\overline{AP}$, $\overline{P'Q}$, and $\overline{Q'B}$ (where P and P' are adjacent integral points on $\overline{AB}$, as are Q and Q'), each of whose integral points is covered by one of the translates $S'_{k+1} + \mathbf{v}$ with $\mathbf{v} \succ \mathbf{0}$. Let l_1, l_2, l_3 be the numbers of integral points on these segments; then $l_1 + l_2 + l_3 = k + 2$.

If $S'_{k+1} + \mathbf{v}$ covers the integral points on $\overline{AP}$, then it covers all of the integral points in the isosceles right triangle that lies below $\overline{AP}$ and has $\overline{AP}$ as its hypotenuse. In fact, all of these points other than those on $\overline{AP}$ itself are covered by $S'_k + \mathbf{v}$; there are $(l_1^2 - l_1)/2$ such points and all are in S'_k. Similarly, the segments $\overline{P'Q}$ and $\overline{Q'B}$ give $(l_2^2 - l_2)/2 + (l_3^2 - l_3)/2$ more points of S'_k which are covered by translates $S'_k + \mathbf{v}$ with $\mathbf{v} \succ \mathbf{0}$.

As in the proof of Lemma 3.1, let T be the set of points in S'_k that are not in $S'_k + \mathbf{v}$ for any $\mathbf{v} \succ \mathbf{0}$; then $|T| = |Z^2 : L|$. The calculations above show that

$$|T| \le |S'_k| + \frac{l_1 + l_2 + l_3}{2} - \frac{l_1^2 + l_2^2 + l_3^2}{2}.$$

Here $|S'_k| = (k+1)(k+2)/2$, and $l_1 + l_2 + l_3$ is just $k + 2$. Given $l_1 + l_2 + l_3$, we minimize $l_1^2 + l_2^2 + l_3^2$ by making the numbers l_1, l_2, l_3 as close to equal as possible; in this case, this means that the minimum occurs when two of them are a and the third is b. Therefore,

$$|T| \le \frac{(k+1)(k+2)}{2} + \frac{k+2}{2} - \frac{2a^2 + b^2}{2},$$

which simplifies to $|T| \le \lfloor (k+2)^2/3 \rfloor$, as desired. □

Corollary 6.3 (due mainly to Wong and Coppersmith [24]). *The largest possible size for the directed Cayley graph of an Abelian group on two generators having diameter k is $\lfloor (k+2)^2/3 \rfloor$.*

Again it is hard to be historically accurate here because different authors have presented results in quite different ways; see the Bermond–Comellas–Hsu survey [3] for more information and references.

Let a, b, and T_k be as in the proof of Theorem 6.2; then the set T_k gives a suitable layout for a network realizing this Cayley graph. In addition to the mesh connections from (i, j) to $(i+1, j)$ and $(i, j+1)$ within T_k, one will also need wraparound connections from $(a+b-1, j)$ to $(0, j+b)$ for $j < a$, from $(a-1, j+a)$ to $(0, j)$ for $j < b$, from $(i, a+b-1)$ to $(i+b, 0)$ for $i < a$, and from $(i+a, a-1)$ to $(i, 0)$ for $i < b$.

In the case $a = b$, one can use an alternate layout in the form of a $3a \times a$ rectangle, with wraparound connections from $(3a, j)$ to $(1, j)$ and from (i, a) to $((i+a) \bmod a, 1)$. This is just a variant of the twisted toroidal mesh, where the long dimension is twisted by a factor of 1/3 rather than 1/2; it is convenient for construction and for embedding a rectangular grid without boundary connections into the network (although this is not particularly useful in the directed case). If $a \ne b$, then one gets a rectangle with some missing nodes or extra nodes along part of one edge, and the cross-connections are slightly messier.

If $k \equiv 1 \pmod 3$, so that $a = b = (k+2)/3$, then one can see from the proof of Theorem 6.2 that the lattice L with generating vectors (a, a) and $(a+b, -b)$ is the unique lattice attaining the bound in the theorem, and hence the Cayley graph attaining the bound in Corollary 6.3 is also unique. However, if $k \not\equiv 1 \pmod 3$, so that a and b differ by 1, then there are two more lattices attaining the bound: the lattice $\tilde{L}$ with generating vectors (a, b) and $(2a, -a)$, and the mirror image with generating vectors (b, a) and $(-a, 2a)$. The latter two give Cayley graphs that are isomorphic to each other, but not to the Cayley graph of $\mathbf{Z}^2/L$ (if $k > 1$), because the Cayley graph of $\mathbf{Z}^2/L$ has cycles of length $2a$ while that of $\mathbf{Z}^2/\tilde{L}$ does not. Therefore, if $k > 0$ and $k \not\equiv 1 \pmod 3$, then there are exactly two Cayley graphs meeting the bound of Corollary 6.3.

If $k > 1$ and $k \equiv 1 \pmod 3$, then the optimal group $\mathbf{Z}^2/L$ is not cyclic; it is isomorphic to $\mathbf{Z}_{3a} \times \mathbf{Z}_a$ by an isomorphism sending $\mathbf{e}_1$ and $\mathbf{e}_2$ to $(1, 0)$ and $(3a-1, 1)$. On the other hand, if $k \not\equiv 1 \pmod 3$, then $(2a + b, a - b)$ is in L and $a - b = \pm 1$, so $\mathbf{e}_2$ is a multiple of $\mathbf{e}_1$ in $\mathbf{Z}^2/L$, and hence $\mathbf{Z}^2/L$ is cyclic. Similarly, $\mathbf{Z}^2/\tilde{L}$ is cyclic since $(3a, b - a) \in \tilde{L}$. One can get the corresponding Cayley graphs directly from the cyclic group $\mathbf{Z}_{\lfloor (k+2)^2/3 \rfloor}$ by using the generator pairs $\{1, (2a + b)/(b - a)\}$ and $\{1, 3a/(a - b)\}$, respectively.

7. Undirected Cayley graphs on three generators. For $d = 3$, we must consider three-dimensional lattice tilings by the regular octahedron $\bar{S}_k$ and its discrete approximation S_k. These shapes do not tile space perfectly, and the best possible efficiency for a lattice covering of $\mathbf{R}^3$ by $\bar{S}_k$ appears to be still open (although there is a good guess, as we shall see). So we will apply our results in reverse, using computed results about the degree-diameter problem to obtain information about lattice tilings by octahedra.

The best three-dimensional toroidal mesh with diameter k is $\mathbf{Z}_{2b_0+1} \times \mathbf{Z}_{2b_1+1} \times \mathbf{Z}_{2b_2+1}$, where $b_i = \lfloor (k + i)/3 \rfloor$; this has about $(8/27)k^3$ vertices for large k. This corresponds to the covering of $\mathbf{R}^3$ by $\bar{S}_1$ using the cubic lattice $(2/3)\mathbf{Z}^3$; this covering has efficiency 2/9.

It turns out that a good lattice to use for coverings with regular octahedra is the body-centered cubic lattice, defined most simply as the set L_{bcc} of points in $\mathbf{Z}^3$ whose coordinates are all odd or all even. If $\mathbf{x}$ is an arbitrary point of $\mathbf{R}^3$, then $\mathbf{x}$ lies in or on one of the unit cubes with vertices in $\mathbf{Z}^3$; two opposite corners of this cube will be in L_{bcc}, say $\mathbf{v}$ and $\mathbf{w}$. Then each coordinate of $\mathbf{x}$ lies between (inclusively) the corresponding coordinates of $\mathbf{v}$ and $\mathbf{w}$, so, letting δ be the l^1 metric on $\mathbf{R}^3$, we have $\delta(\mathbf{v}, \mathbf{x}) + \delta(\mathbf{x}, \mathbf{w}) = \delta(\mathbf{v}, \mathbf{w}) = 3$. Hence, either $\delta(\mathbf{v}, \mathbf{x}) \le 3/2$ or $\delta(\mathbf{w}, \mathbf{x}) \le 3/2$. This shows that $|\bar{S}_{3/2}| + L_{\text{bcc}} = \mathbf{R}^3$. Now, L_{bcc} has generators $(2, 0, 0)$, $(0, 2, 0)$, and $(1, 1, 1)$, giving a matrix with determinant 4, while the volume of $\bar{S}_{3/2}$ is 9/2, so this lattice covering of $\mathbf{R}^3$ has efficiency 8/9. A fundamental region for the lattice can be obtained by truncating each of the corners of the octahedron, giving an Archimedean solid whose faces are eight regular hexagons and six squares.

The same reasoning shows that, for $k \ge 1$, one can get a lattice covering of $\mathbf{Z}^3$ by S_k using the slightly distorted body-centered cubic lattice $L_{\text{bcc}}(a_1, a_2, a_3)$ with generating vectors $(2a_1, 0, 0)$, $(0, 2a_2, 0)$, and (a_1, a_2, a_3), where $a_i = \lfloor (2k + i)/3 \rfloor$. This gives a Cayley graph of size $4a_1a_2a_3$, or approximately $(32/27)k^3$ for large k. This is an improvement over the best toroidal mesh of diameter k; it is about four times as good for large k.

One can lay out the Cayley graph for $\mathbf{Z}^3/L_{\text{bcc}}(a_1, a_2, a_3)$ in the form of a $2a_1 \times 2a_2 \times a_3$ mesh. Opposite $2a_i \times a_3$ sides are connected to each other as in the usual toroidal mesh, but the toroidal connections between the top and bottom $2a_1 \times 2a_2$ sides are twisted in two directions: node (j_1, j_2, a_3) is connected to node $(j_1 \pm a_1, j_2 \pm a_2, 1)$, where the signs are chosen to give numbers between 1 and $2a_i$, inclusive. Routing algorithms and embeddings of rectangular grids work here just as they did in the two-dimensional version.

Two questions now arise. First, can one improve the efficiency by making small adjustments to the discrete lattice, as we did in the two-generator cases? Second, can one get better results by using a completely different lattice? The answers to these questions are not immediately clear, so we will approach the problem from another direction.

One can write computer programs to examine various groups, choose all (or at least many) possible sets of a certain number of generators for the group, and compute the resulting diameters. Dinneen has performed many such computations, some using exhaustive search of generator sets and others using random sampling, on a number of different kinds of groups, resulting in new best-known graphs for the degree-diameter problem; see, for instance, Dinneen and Hafner [10]. Some of Dinneen's earlier unpublished computations were for Abelian (usually cyclic) groups of diameter up to 10 on various numbers of generators.

The authors have written a program to extend these calculations. The program does an exhaustive search of generating sets for each Abelian group but avoids examining many generating sets which give Cayley graphs that are isomorphic to ones already examined; for instance, in the case of a cyclic group $\mathbf{Z}_n$, one may assume that the first generator is a divisor of n. Here "exhaustive search" means that all Abelian groups of size up to $|S_k| = (4k^3 + 6k^2 + 8k + 3)/3$ were examined, so the results definitely give the largest possible Cayley graph of an Abelian group with diameter k. The program uses bit manipulations adapted from (but simpler than) those of Dougherty and Janwa [11], which gave algorithms for diameter computations for Cayley graphs of Abelian groups of exponent 2.

It turns out that, for each k for which the calculation has been done so far (up to 18), the best Abelian Cayley graph has been obtained from a cyclic group. The results of the computation are shown in Table 7.1. This extends (and corrects an erroneous final entry in) a similar table given by Chen and Jia [5].

The first column is the desired diameter k. The second column gives the largest size one could hope for of an undirected Cayley graph of an Abelian group on three generators. The next two columns give the sizes attained by the best possible ordinary toroidal mesh and the twisted toroidal mesh described above. Next comes n_c, the computed largest n such that $\mathbf{Z}_n$ has three generators giving it an undirected diameter of k. Then comes a triple of generators of $\mathbf{Z}_{n_c}$ attaining this diameter (this is not always unique, but only one generator set is given here). The final two columns give the efficiencies of the corresponding lattice coverings of $\mathbf{Z}^3$ by S_k and of $\mathbf{R}^3$ by $\bar{S}_{k+3/2}$ (see Proposition 4.1).

Some interesting observations can be made from Table 7.1. First, note that the twisted toroidal meshes do almost as well as the optimal cyclic groups. Also note that the numbers in the second-to-last column do seem to be getting close to 8/9 for larger k; this provides evidence that the body-centered cubic lattice gives the best lattice covering of $\mathbf{R}^3$ by S_1.

One can confirm this more strongly by reconstructing the lattices L for which $\mathbf{Z}^3/L$ gives these optimal cyclic groups. For instance, look at $k = 10$, for which we have the cyclic group $\mathbf{Z}_{1393}$ with generating set $\{1, 92, 106\}$. There is a unique homomorphism from $\mathbf{Z}^3$ to $\mathbf{Z}_{1393}$ which sends $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ to $1, 92, 106$, and the desired lattice L is just the kernel of this homomorphism; this means that

$$L = \{(x_1, x_2, x_3) \in \mathbf{Z}^3 : x_1 + 92x_2 + 106x_3 \equiv 0 \pmod{1393}\}.$$

One can easily find three vectors in L, namely $(1393, 0, 0)$, $(92, -1, 0)$, and $(106, 0, -1)$;

TABLE 7.1
Best undirected Cayley graphs of cyclic groups, three generators.

k	$\lvert S_k\rvert$	Toroidal	Twisted	n_c	Generators	$n_c/\lvert S_k\rvert$	$n_c/\operatorname{vol}(\bar{S}_{k+3/2})$
0	1	1		1		1	.222222
1	7	3	4	7	1,2,3	1	.336000
2	25	9	16	21	1,2,8	.840000	.367347
3	63	27	48	55	1,5,21	.873016	.452675
4	129	45	108	117	1,16,22	.906977	.527423
5	231	75	192	203	1,7,57	.878788	.554392
6	377	125	320	333	1,9,73	.883289	.592000
7	575	175	500	515	1,46,56	.895652	.628944
8	833	245	720	737	1,11,133	.884754	.644700
9	1159	343	1008	1027	1,13,157	.886109	.665371
10	1561	441	1372	1393	1,92,106	.892377	.686940
11	2047	567	1792	1815	1,15,241	.886663	.696960
12	2625	729	2304	2329	1,17,273	.887238	.709953
13	3303	891	2916	2943	1,154,172	.891008	.724015
14	4089	1089	3600	3629	1,19,381	.887503	.730892
15	4991	1331	4400	4431	1,21,421	.887798	.739795
16	6017	1573	5324	5357	1,232,254	.890311	.749668
17	7175	1859	6336	6371	1,23,553	.887944	.754664
18	8473	2197	7488	7525	1,25,601	.888115	.761139

the matrix with these three vectors as rows has determinant $1393 = |Z^3 : L|$, so these vectors generate L. Now one can perform elementary operations to reduce these vectors to a smaller set of generators for L, such as $(7,7,7)$, $(8,-7,6)$, and $(6,8,-7)$. These vectors are quite close to the vectors $(7,7,7)$, $(7,-7,7)$, and $(7,7,-7)$, which generate a scaled-up body-centered cubic lattice (in fact, the latter lattice gives the twisted toroidal mesh of size 1372 mentioned in the table). Similarly, one finds that the other lattices corresponding to the generators in Table 7.1 are almost body-centered cubic.

There are definite patterns in Table 7.1: every third k gives groups and generators of the same form. These patterns can be generalized, giving the following result.

THEOREM 7.1. *For all $k \geq 0$, there is an undirected Cayley graph on three generators of an Abelian (in fact, cyclic) group which has diameter k and size n, where*

$$n = \begin{cases} (32k^3 + 48k^2 + 54k + 27)/27 & \textit{if } k \equiv 0 \pmod 3, \\ (32k^3 + 48k^2 + 78k + 31)/27 & \textit{if } k \equiv 1 \pmod 3, \\ (32k^3 + 48k^2 + 54k + 11)/27 & \textit{if } k \equiv 2 \pmod 3. \end{cases}$$

Proof. We will show the existence of lattices $L_k \subseteq \mathbf{Z}^3$ such that $\mathbf{Z}^3/L_k$ is cyclic, $S_k + L_k = \mathbf{Z}^3$, and $|\mathbf{Z}^3 : L|$ is the n specified in the theorem.

Let $a = \lceil 2k/3 \rceil$. For each k, we define L_k by specifying three generating vectors

$\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ for it, as follows:

$$\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 = \begin{cases} (a{+}1, a, a), (a, -a, a{+}1), (a{+}1, a{-}1, -a{-}1) & \text{if } k \equiv 0 \pmod 3, \\ (a, a, a), (a{+}1, -a, a{-}1), (a{-}1, a{+}1, -a) & \text{if } k \equiv 1 \pmod 3, \\ (a, a, a{-}1), (a{-}1, -a, a), (a, a{-}1, -a) & \text{if } k \equiv 2 \pmod 3. \end{cases}$$

A simple determinant computation shows that $|\mathbf{Z}^3 : L_k|$ is $(2a^2 + a + 1)(2a + 1)$, $4a^3 + 3a$, or $(2a^2 - a + 1)(2a - 1)$ in the respective case $k \equiv 0$, $k \equiv 1$, or $k \equiv 2 \pmod 3$. Since a is, respectively, $2k/3$, $(2k + 1)/3$, or $(2k + 2)/3$, the index $|\mathbf{Z}^3 : L_k|$ works out to be the desired value n.

For $k \equiv 0 \pmod 3$, the following vectors are in L_k:

$$\begin{aligned} \mathbf{v}_2 + \mathbf{v}_3 &= (2a{+}1, -1, 0), \\ \mathbf{v}_1 + (2a{-}1)\mathbf{v}_2 + 2a\mathbf{v}_3 &= (4a^2{+}2a{+}1, 0, -1). \end{aligned}$$

Hence, we have $\mathbf{e}_2 = (2a{+}1)\mathbf{e}_1$ and $\mathbf{e}_3 = (4a^2{+}2a{+}1)\mathbf{e}_1$ in $\mathbf{Z}^3/L_k$, so $\mathbf{e}_1$ generates $\mathbf{Z}^3/L_k$. Thus $\mathbf{Z}^3/L_k$ is isomorphic to $\mathbf{Z}_n$ via an isomorphism taking $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ to $1, 2a{+}1, 4a^2{+}2a{+}1$. Similarly, for $k \equiv 1 \pmod 3$ we have

$$\begin{aligned} a\mathbf{v}_2 + (a{-}1)\mathbf{v}_3 &= (2a^2{-}a{+}1, -1, 0), \\ (a{+}1)\mathbf{v}_2 + a\mathbf{v}_3 &= (2a^2{+}a{+}1, 0, -1), \end{aligned}$$

so $\mathbf{Z}^3/L_k$ is isomorphic to $\mathbf{Z}_n$ with generators $1, 2a^2{-}a{+}1, 2a^2{+}a{+}1$, and for $k \equiv 2 \pmod 3$ we have

$$\begin{aligned} \mathbf{v}_2 + \mathbf{v}_3 &= (2a{-}1, -1, 0), \\ \mathbf{v}_1 + (2a{-}1)\mathbf{v}_2 + 2a\mathbf{v}_3 &= (4a^2{-}2a{+}1, 0, -1), \end{aligned}$$

so $\mathbf{Z}^3/L_k$ is isomorphic to $\mathbf{Z}_n$ with generators $1, 2a{-}1, 4a^2{-}2a{+}1$.

It remains to show that $S_k + L_k = \mathbf{Z}^3$. We will do only the case $k \equiv 1 \pmod 3$ here; the other two cases are handled by the same method, but with a few more subcases because of less symmetry.

For $k = 1$ one just has to show that $\mathbf{Z}_7$ with generators $1, 2, 4$ has diameter 1; this is trivial to do directly, so we may assume $k > 1$ and hence $a > 1$.

Let $\mathbf{v}_4 = \mathbf{v}_1 - \mathbf{v}_2 - \mathbf{v}_3 = (-a, a{-}1, a{+}1)$. Then the vectors $\pm\mathbf{v}_i$ for $i = 1, 2, 3, 4$ give one member of L_k strictly within each of the eight octants of $\mathbf{Z}^3$, and all of the coordinates of these vectors have absolute value at most $a{+}1$.

We must show that each $\mathbf{x} \in \mathbf{Z}^3$ is in $S_k + L_k$. This is equivalent to showing that there is a member $\mathbf{w}$ of L_k such that $\mathbf{x} - \mathbf{w} \in S_k$, which in turn is equivalent to $\delta(\mathbf{x}, \mathbf{w}) \le k$, where δ is the l^1 (Manhattan) metric on $\mathbf{Z}^3$. Note that if $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are such that each coordinate of $\mathbf{y}$ is between (inclusively) the corresponding coordinates of $\mathbf{x}$ and $\mathbf{z}$, then $\delta(\mathbf{x}, \mathbf{y}) + \delta(\mathbf{y}, \mathbf{z}) = \delta(\mathbf{x}, \mathbf{z})$. From now on, we will state this situation more briefly as "$\mathbf{y}$ lies between $\mathbf{x}$ and $\mathbf{z}$."

Suppose we are given $\mathbf{x} \in \mathbf{Z}^3$. The idea is to repeatedly reduce $\mathbf{x}$ by adding members of L_k to it until one reaches a vector which is within l^1-distance k of $\mathbf{0}$ or some other known member of L_k.

First we will reduce $\mathbf{x}$ to a vector whose coordinates all have absolute value at most $a{+}1$. Suppose $\mathbf{x}$ does not already have this property. Let $\mathbf{v}$ be one of the vectors $\pm\mathbf{v}_i$ $(i = 1, 2, 3, 4)$ such that the coordinates of $\mathbf{v}$ have the same signs as the corresponding coordinates of $\mathbf{x}$; if a coordinate of $\mathbf{x}$ is 0, then either sign is allowed for the corresponding coordinate of $\mathbf{v}$. Now look at $\mathbf{x}' = \mathbf{x} - \mathbf{v}$. If a coordinate of $\mathbf{x}$ has absolute value $\leq a{+}1$, then the corresponding coordinate of $\mathbf{x}'$ will also have absolute value $\leq a{+}1$ because of the sign matching and the fact that the coordinates of $\mathbf{v}$ have absolute value $\leq a{+}1$. If a coordinate of $\mathbf{x}$ has absolute value $> a{+}1$, then the corresponding coordinate of $\mathbf{x}'$ will be strictly smaller in absolute value. Therefore, repeating this procedure will lead, after finitely many steps, to a vector whose coordinates all have absolute value at most $a{+}1$.

If this new $\mathbf{x}$ lies between $\mathbf{0}$ and one of the vectors $\pm\mathbf{v}_i$, then we have $\delta(\mathbf{0}, \mathbf{x}) + \delta(\mathbf{x}, \pm\mathbf{v}_i) = \delta(\mathbf{0}, \pm\mathbf{v}_i)$. But all of the vectors $\pm\mathbf{v}_i$ satisfy $\delta(\mathbf{0}, \pm\mathbf{v}_i) = 2k + 1$; since $\delta(\mathbf{0}, \mathbf{x})$ and $\delta(\mathbf{x}, \pm\mathbf{v}_i)$ are both integers, one of them must be at most k, so we are done with this $\mathbf{x}$.

We now break into cases depending on which octant the new $\mathbf{x}$ lies in. Since L_k is centrosymmetric, we need only handle the octants containing $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, and $\mathbf{v}_4$. Also, L_k is invariant under cyclic permutations of the three coordinates since these permutations leave $\mathbf{v}_1$ fixed and permute $\mathbf{v}_2$, $\mathbf{v}_3$, $\mathbf{v}_4$; hence, we may assume that the new $\mathbf{x}$ is in the octant of $\mathbf{v}_1$ or the octant of $\mathbf{v}_2$.

First, suppose that $\mathbf{x}$ is now in the octant of $\mathbf{v}_1$ (where all three coordinates are nonnegative). If $\mathbf{x}$ is between $\mathbf{0}$ and $\mathbf{v}_1$, we are done. If two or more of the coordinates of $\mathbf{x}$ are equal to $a{+}1$, say (by cyclic symmetry of L_k) $\mathbf{x} = (a{+}1, a{+}1, r)$, then we have $\delta(\mathbf{x}, \mathbf{v}_1) \leq k$, unless $k = 4$ and $r = 0$, in which case $\delta(\mathbf{x}, \mathbf{v}_1 + \mathbf{v}_3) = k$.

If $\mathbf{x}$ has exactly one coordinate equal to $a{+}1$, say $\mathbf{x} = (a{+}1, r, s)$ with $0 \leq r{,}s \leq a$, then we can subtract $\mathbf{v}_1$ from $\mathbf{x}$ to get $\mathbf{x}' = (1, r{-}a, s{-}a)$, which is in the octant containing $-\mathbf{v}_4$. If $\mathbf{x}'$ lies between $\mathbf{0}$ and $-\mathbf{v}_4$, we are done. If not, then $r = 0$. Now let $\mathbf{x}'' = \mathbf{x}' + \mathbf{v}_4 = (-a{+}1, -1, s{+}1)$, which lies between $\mathbf{0}$ and $-\mathbf{v}_3$ unless $s = a$, in which case $\mathbf{x} = (a{+}1, 0, a)$ and $\delta(\mathbf{x}, \mathbf{v}_2) = a{+}1 \leq k$.

The procedure in the case where $\mathbf{x}$ is in the octant of $\mathbf{v}_2$ is similar. Either $\mathbf{x} = (r, s, t)$ lies between $\mathbf{0}$ and $\mathbf{v}_2$, or $s = -a{-}1$, or $t \geq a$. In the latter cases, let $\mathbf{x}' = \mathbf{x} - \mathbf{v}_2$. When $s = -a{-}1$, we have $\mathbf{x}' = (r{-}a{-}1, -1, t{-}a{+}1)$; either this lies between $\mathbf{0}$ and one of the vectors $\pm\mathbf{v}_i$, or $\mathbf{x} + \mathbf{v}_3$ does. When $s \geq -a$ but $t \geq a$, try $\mathbf{x}' - \mathbf{v}_4$; it either lies between $\mathbf{0}$ and some $\pm\mathbf{v}_i$ or is $(-1, 1, -a)$, $(a, 1, -a)$, or $(a, 1, -a{+}1)$. These last three lie within δ-distance k of $\mathbf{v}_3 - \mathbf{v}_1$, $\mathbf{v}_3$, and either $\mathbf{v}_3 + \mathbf{v}_2 - \mathbf{v}_1$ or $\mathbf{v}_3 + \mathbf{v}_2$, respectively. □

The authors conjecture that the graphs given by this theorem are actually the largest undirected Cayley graphs of Abelian groups on three generators for each diameter k.

This conjecture would imply that the lattice covering of $\mathbf{R}^3$ by $\bar{S}_{3/2}$ using the lattice L_{bcc} is optimal; that is, $8/9$ is the best possible efficiency for a lattice covering by regular octahedra. The latter statement seems quite plausible but remains unproved at this point. However, we can prove the partial result that a "small" adjustment to L_{bcc} cannot improve the covering. See the following theorem.

THEOREM 7.2. *Among those lattices L for which $\bar{S}_{3/2} + L = \mathbf{R}^3$, the lattice L_{bcc} is locally optimal; that is, for any other lattice L sufficiently near L_{bcc} such that $\bar{S}_{3/2} + L = \mathbf{R}^3$, the efficiency of the covering using L is less than $8/9$.*

Proof. Let us use the vectors $\mathbf{v}_1 = (1, 1, 1)$, $\mathbf{v}_2 = (1, -1, 1)$, and $\mathbf{v}_3 = (1, 1, -1)$ as generating vectors for L_{bcc}; then a nearby lattice L will be generated by nearby vectors $\mathbf{v}'_1 = (a_1, b_1, c_1)$, $\mathbf{v}'_2 = (a_2, b_2, c_2)$, and $\mathbf{v}'_3 = (a_3, b_3, c_3)$. We can concatenate the three vectors $\mathbf{v}'_1$, $\mathbf{v}'_2$, $\mathbf{v}'_3$ to get a single vector $\mathbf{v}'$ in $\mathbf{R}^9$; similarly, let $\mathbf{v}$ be the concatenation $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$. Let $F(\mathbf{v}')$ be the determinant of the matrix with rows $\mathbf{v}'_1$, $\mathbf{v}'_2$, $\mathbf{v}'_3$. Note that $F(\mathbf{v}) = 4$; we must see that this point is a strict local maximum of

$F(\mathbf{v}')$ for those points $\mathbf{v}'$ satisfying the constraint that $\bar{S}_{3/2} + L = \mathbf{R}^3$. We compute the gradient of F at the point $\mathbf{v}$ to be $\mathbf{g} = (0, 2, 2, 2, -2, 0, 2, 0, -2)$.

Using the lattice L_{bcc}, the point $(1/2, 1/2, 1/2)$, in the center of a face of $\bar{S}_{3/2}$, is covered by only two copies of $\bar{S}_{3/2}$, namely $\bar{S}_{3/2}$ itself and $\bar{S}_{3/2} + \mathbf{v}_1$, and it is on the boundary (a face) of each of these copies. If the lattice is altered slightly so that these two copies no longer touch, then the points in between will not be covered by any copy. In particular, if L is near L_{bcc} but $a_1 + b_1 + c_1 > 3$, then the point $(1/2, 1/2, 1/2 + \varepsilon)$ for small positive ε will not be in $\bar{S}_{3/2} + L$. So the constraint $\bar{S}_{3/2} + L = \mathbf{R}^3$ gives us the linear inequality $a_1 + b_1 + c_1 \leq 3$. We will rewrite this as

$$\mathbf{u}_1 \cdot \mathbf{v}' \leq 3, \quad \text{where } \mathbf{u}_1 = (1, 1, 1, 0, 0, 0, 0, 0, 0).$$

The same argument for points on the other faces of the octahedron gives inequalities

$$\begin{aligned}
\mathbf{u}_2 \cdot \mathbf{v}' &\leq 3, \quad \text{where } \mathbf{u}_2 = (0, 0, 0, 1, -1, 1, 0, 0, 0),\\
\mathbf{u}_3 \cdot \mathbf{v}' &\leq 3, \quad \text{where } \mathbf{u}_3 = (0, 0, 0, 0, 0, 0, 1, 1, -1),\\
\mathbf{u}_4 \cdot \mathbf{v}' &\leq 3, \quad \text{where } \mathbf{u}_4 = (-1, 1, 1, 1, -1, -1, 1, -1, -1).
\end{aligned}$$

Next, consider the point $(1, 0, 1/2)$. This is in $\bar{S}_{3/2} + \mathbf{y}$ for four members $\mathbf{y}$ of L_{bcc}, namely $\mathbf{0}$, $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_2 + \mathbf{v}_3$, and it is an edge point of each of these four copies. If the lattice is altered slightly, then a gap can open up near this point even if there are no gaps between octahedra adjacent at a face as above.

Specifically, if $\mathbf{v}'$ is near $\mathbf{v}$, ε is a very small positive number, and we define the point $\mathbf{x}$ by the linear equations

$$\begin{aligned}
\mathbf{x} \cdot (1, -1, -1) &= \mathbf{v}'_1 \cdot (1, -1, -1) + 3/2 + \varepsilon,\\
\mathbf{x} \cdot (-1, -1, 1) &= (\mathbf{v}'_2 + \mathbf{v}'_3) \cdot (-1, -1, 1) + 3/2 + \varepsilon,\\
\mathbf{x} \cdot (-1, 1, -1) &= \mathbf{v}'_2 \cdot (-1, 1, -1) + 3/2 + \varepsilon,
\end{aligned}$$

then $\mathbf{x}$ will be a point near $(1, 0, 1/2)$ which is not in $\bar{S}_{3/2} + \mathbf{y}$ for $\mathbf{y} \in \{\mathbf{v}'_1, \mathbf{v}'_2, \mathbf{v}'_2 + \mathbf{v}'_3\}$. Adding up the three given equations yields

$$\mathbf{x} \cdot (-1, -1, -1) = \mathbf{v}'_1 \cdot (1, -1, -1) + \mathbf{v}'_2 \cdot (-2, 0, 0) + \mathbf{v}'_3 \cdot (-1, -1, 1) + 9/2 + 3\varepsilon.$$

If the right-hand side of this equation is less than $-3/2$, then $\mathbf{x}$ will not be in $\bar{S}_{3/2}$ either, and hence will not be in $\bar{S}_{3/2} + L$. Since ε can be arbitrarily small, in order to have $\bar{S}_{3/2} + L = \mathbf{R}^3$, it is necessary to have

$$\mathbf{v}'_1 \cdot (1, -1, -1) + \mathbf{v}'_2 \cdot (-2, 0, 0) + \mathbf{v}'_3 \cdot (-1, -1, 1) \geq -6.$$

This can be rewritten as

$$\mathbf{u}_5 \cdot \mathbf{v}' \leq 6, \quad \text{where } \mathbf{u}_5 = (-1, 1, 1, 2, 0, 0, 1, 1, -1).$$

The same argument can be performed using the octahedra around $(1, 0, 1/2)$ in the opposite order, and there are 23 other points on the edges of $\bar{S}_{3/2}$ where the same configuration occurs. But one only gets six distinct inequalities from this; the other five are

$$\begin{aligned}
\mathbf{u}_6 \cdot \mathbf{v}' &\leq 6, \quad \text{where } \mathbf{u}_6 = (0, 2, 0, 1, -1, 1, 1, -1, -1),\\
\mathbf{u}_7 \cdot \mathbf{v}' &\leq 6, \quad \text{where } \mathbf{u}_7 = (-1, 1, 1, 1, -1, 1, 2, 0, 0),\\
\mathbf{u}_8 \cdot \mathbf{v}' &\leq 6, \quad \text{where } \mathbf{u}_8 = (1, 1, 1, 0, -2, 0, 1, -1, -1),\\
\mathbf{u}_9 \cdot \mathbf{v}' &\leq 6, \quad \text{where } \mathbf{u}_9 = (0, 0, 2, 1, -1, -1, 1, 1, -1),\\
\mathbf{u}_{10} \cdot \mathbf{v}' &\leq 6, \quad \text{where } \mathbf{u}_{10} = (1, 1, 1, 1, -1, -1, 0, 0, -2).
\end{aligned}$$

Note that all 10 of these inequalities are satisfied with equality when $\mathbf{v}' = \mathbf{v}$. Hence, they can be rewritten as $\mathbf{u}_i \cdot (\mathbf{v}' - \mathbf{v}) \leq 0$ for $i = 1, 2, \ldots, 10$.

One can easily check that the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_7$ are linearly independent; their common null space (i.e., the set of $\mathbf{w}$ such that $\mathbf{u}_i \cdot \mathbf{w} = 0$ for all $i \leq 7$) is generated by the independent vectors $\mathbf{w}_1 = (1, 0, -1, 1, 0, -1, 1, 0, 1)$ and $\mathbf{w}_2 = (-1, 1, 0, -1, -1, 0, -1, 1, 0)$. Also, we have

$$\mathbf{g} = \mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3 + \mathbf{u}_4 = \mathbf{u}_5 + \mathbf{u}_8 = \mathbf{u}_6 + \mathbf{u}_9 = \mathbf{u}_7 + \mathbf{u}_{10}.$$

Let C be the closed cone consisting of all vectors $\mathbf{t}$ in the subspace spanned by $\mathbf{u}_1, \ldots, \mathbf{u}_7$ such that $\mathbf{u}_i \cdot \mathbf{t} \leq 0$ for all $i \leq 10$. Then the above equations imply that $\mathbf{g} \cdot \mathbf{t} \leq 0$ for all $\mathbf{t}$ in C, and equality can hold only when $\mathbf{t} = \mathbf{0}$. In particular, we have $\mathbf{g} \cdot \mathbf{t}_0 < 0$ for any unit vector $\mathbf{t}_0$ in C. The set of such $\mathbf{t}_0$ is closed and bounded, hence compact, so there is a positive number ε such that $\mathbf{g} \cdot \mathbf{t}_0 < -\varepsilon$ for all such $\mathbf{t}_0$. It follows that there is a neighborhood U of $\mathbf{g}$ such that, for any $\mathbf{g}'$ in U and any unit vector $\mathbf{t}_0$ in C, $\mathbf{g}' \cdot \mathbf{t}_0 < 0$. Since C is a cone, we have $\mathbf{g}' \cdot \mathbf{t} < 0$ for all $\mathbf{g}' \in U$ and all nonzero $\mathbf{t} \in C$.

We can compute that, for any real numbers r and s, the determinant for the lattice given by $\mathbf{v} + r\mathbf{w}_1 + s\mathbf{w}_2$ is

$$F(\mathbf{v} + r\mathbf{w}_1 + s\mathbf{w}_2) = 4(1 - r)(1 + s)(1 + r - s).$$

If $|r| + |s| < 1$, then $1 - r$, $1 + s$, and $1 + r - s$ are positive numbers with arithmetic mean 1, so their geometric mean is at most 1; this means that $F(\mathbf{v} + r\mathbf{w}_1 + s\mathbf{w}_2) \leq 4$. Equality holds only when the above three numbers are equal, which is when $r = s = 0$.

Let U' be a convex neighborhood of $\mathbf{v}$ so small that $(\operatorname{grad} F)(\mathbf{v}') \in U$ for all $\mathbf{v}' \in U'$. Now, any vector $\mathbf{v}'$ sufficiently close to $\mathbf{v}$ can be expressed as $\mathbf{v} + \mathbf{t}_1 + \mathbf{t}_2$, so $\mathbf{t}_1$ is a (small) linear combination of $\mathbf{w}_1$ and $\mathbf{w}_2$, $\mathbf{t}_2$ is a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_7$, and both $\mathbf{v} + \mathbf{t}_1$ and $\mathbf{v} + \mathbf{t}_1 + \mathbf{t}_2$ are in U'. If $\mathbf{v}'$ satisfies the condition $\bar{S}_{3/2} + L = \mathbf{R}^3$ and is near $\mathbf{v}$, then we must have $\mathbf{u}_i \cdot (\mathbf{v}' - \mathbf{v}) \leq 0$ for all $i \leq 10$, so $\mathbf{u}_i \cdot \mathbf{t}_2 \leq 0$ for all $i \leq 10$ (since $\mathbf{u}_i \cdot \mathbf{t}_1 = 0$), so $\mathbf{t}_2 \in C$. We have $F(\mathbf{v} + \mathbf{t}_1) \leq 4$, with equality holding only when $\mathbf{t}_1 = \mathbf{0}$. If $\mathbf{t}_2$ is nonzero, then for any $\mathbf{t}$ on the segment from $\mathbf{v} + \mathbf{t}_1$ to $\mathbf{v} + \mathbf{t}_1 + \mathbf{t}_2$ we have $\mathbf{t} \in U'$, so $(\operatorname{grad} F)(\mathbf{t}) \in U$, so $(\operatorname{grad} F)(\mathbf{t}) \cdot \mathbf{t}_2 < 0$; it follows that $F(\mathbf{v} + \mathbf{t}_1 + \mathbf{t}_2) < F(\mathbf{v} + \mathbf{t}_1)$. Therefore, $F(\mathbf{v}') \leq F(\mathbf{v})$, with equality holding only when $\mathbf{v}' = \mathbf{v}$. So $\mathbf{v}$ gives a local maximum of F, as desired. □

It is still possible (though very unlikely) that a lattice quite different from L_{bcc} gives a more efficient covering. Theoretically, the search for an optimal lattice can be set up as a large optimization problem and solved once and for all, but this appears to be a formidable task.

One could begin this task by considering an arbitrary lattice L such that $\bar{S}_{3/2} + L$ covers $\mathbf{R}^3$, and this covering is reasonably efficient (at least as efficient as the covering from L_{bcc}). Such a lattice is generated by vectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, and we can carefully choose these generators so as to limit their lengths. In particular, we can choose $\mathbf{v}_1$ to be a nonzero member of L with minimal length. We can then choose $\mathbf{v}_2$ in L whose distance from the subspace of $\mathbf{R}^3$ spanned by $\mathbf{v}_1$ is as small as possible (but nonzero), and adjust $\mathbf{v}_2$ by subtracting an integer multiple of $\mathbf{v}_1$ so as to ensure that the closest integer multiple of $\mathbf{v}_1$ to $\mathbf{v}_2$ is $\mathbf{0}$. One can similarly choose $\mathbf{v}_3$ to be as close as possible to (but not in) the subspace spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$. These three chosen vectors will be a set of generating vectors for L. In order to have $\bar{S}_{3/2} + L = \mathbf{R}^3$, it is necessary that the length of $\mathbf{v}_1$ be no more than the diameter of $\bar{S}_{3/2}$; there are similar but slightly larger bounds on the lengths of $\mathbf{v}_2$ and $\mathbf{v}_3$. This limits our search for $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ to a compact subset of nine-dimensional space. We must find the point in this subset which maximizes $\det(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ subject to the constraint that $\bar{S}_{3/2} + L = \mathbf{R}^3$.

This constraint looks infinitary, but it can actually be reduced to finitely many sets of linear inequalities. To see this, note that, using the above upper bounds on the lengths of the vectors $\mathbf{v}_i$ along with the assumed lower bound on the lattice determinant (the covering must be at least as efficient as that from L_{bcc}), we can get lower bounds on the lengths of the vectors $\mathbf{v}_i$, the angles between them, and associated quantities such as the distance from $\mathbf{v}_3$ to the plane spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$. These will allow us to get upper bounds on the absolute values of integers a_1, a_2, a_3 such that $\bar{S}_{3/2} + a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + a_3\mathbf{v}_3$ overlaps or almost touches $\bar{S}_{3/2}$. (In other words, we get an upper bound on the number M from the proof of Proposition 4.4.) So we only have to consider finitely many of the lattice translates of $\bar{S}_{3/2}$ when trying to cover the space near $\bar{S}_{3/2}$ (which is all that is needed, by Proposition 4.3).

There are only finitely many configurations (specifications of arrangements and overlaps) for these finitely many translates of $\bar{S}_{3/2}$. For each such configuration, the assertion that there are no "gaps" in the coverage of the space near $\bar{S}_{3/2}$ becomes a list of linear inequalities like the inequalities $\mathbf{u}_i \cdot \mathbf{v}' \leq b$ from the proof of Theorem 7.2. So we need to optimize a cubic function (the lattice determinant) subject to a list of linear inequalities in order to find the optimal version of each configuration, and then compare the resulting values to find the best configuration.

Unfortunately, there is a very large number of possible configurations (for an example of the possibilities for complicated configurations, see Figure 8.2 later in this paper), so this finite computation appears to be beyond our reach at present. Of course, a different approach to the problem might lead to a more feasible computation.

One might hope to be able to use the arguments of Proposition 4.4 and Theorem 4.5 in reverse to get an upper bound on the efficiency of lattice coverings of $\mathbf{R}^3$ by the octahedron $\bar{S}_1$ by showing that any extremely efficient real lattice covering would lead to integer lattice coverings more efficient than what the computation actually found. To do this, one would fix a value for the distance ρ from Proposition 4.4 and then use the method described above to get an upper bound on the number M from that proposition. If there is actually a lattice covering of $\mathbf{R}^3$ by $\bar{S}_1$ using the lattice L generated by $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ having a specified large determinant (equivalently, a specified large efficiency), then we can round the coordinates of these vectors to the nearest multiples of $1/k$ to get vectors $\mathbf{v}_1'$, $\mathbf{v}_2'$, $\mathbf{v}_3'$ generating a lattice L'. By Proposition 4.4, if $1/(2k)$ is less than $\eta = \rho/M$, then we will have $\bar{S}_{1+3\rho} + L' = \mathbf{R}^3$, so $\bar{S}_{k+3k\rho} + kL' = \mathbf{R}^3$. But kL' is an integer lattice; if n is its determinant, then this lattice covering will yield an Abelian Cayley graph on three generators with size n and diameter at most $k + 3k\rho$. The fact that L' is close to L means that we can get a lower bound on n from the determinant of L. If the actual computational search showed that there is no Abelian Cayley graph of such a size for this diameter, then our original assumption that there was a lattice L giving a covering of that efficiency must have been false.

Unfortunately, the constants involved are such that even the large computation done so far does not suffice to get a bound less than 1 for the efficiency of L (even if we are optimistic enough to assume that M is as small as 3 or 4). It probably requires searches for values of k larger than 500 in order to get actual results from this method; such searches are completely out of range at the moment.

8. Directed Cayley graphs on three generators. For the directed case of three generators, we want to study lattice coverings of $\mathbf{R}^3$ by the trirectangular tetrahedron $\bar{S}_k'$. (Since lattice covering efficiency is affine invariant, it makes no difference

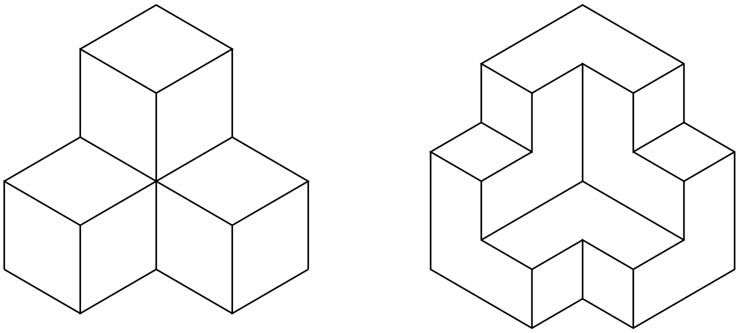


FIG. 8.1. *Two subsets of the tetrahedron $\bar{S}_1'$ which tile space.*

which particular tetrahedron we consider.) One hopes that one can discretize these coverings to get good lattice coverings of $\mathbf{Z}^3$ by S_k', and hence good directed Cayley graphs.

The best three-dimensional directed toroidal mesh with diameter k is $\mathbf{Z}_{b_0+1} \times \mathbf{Z}_{b_1+1} \times \mathbf{Z}_{b_2+1}$, where $b_i = \lfloor (k+i)/3 \rfloor$; this has about $(1/27)k^3$ vertices for large k and corresponds to the covering of $\mathbf{R}^3$ by $\bar{S}_1'$ using the cubic lattice $(1/3)\mathbf{Z}^3$. This covering has efficiency 2/9.

It is more difficult to find a candidate for a good covering lattice (or, equivalently, a large subset which gives a lattice tiling) for the tetrahedron than it was for the octahedron. One possible method is to try to find the three-dimensional analogue of the L-tromino used for the triangle; this leads one to consider the tetracube shown on the left of Figure 8.1. In order for the shape to fit into $\bar{S}_1'$, the edge-length of the subcubes should be 1/4. It is easy to see that this shape does indeed tile space, using the lattice generated by $(1/2, 0, 0)$, $(0, 1/2, 0)$, and $(1/4, 1/4, -1/4)$ (this is just $(1/4)L_{\mathrm{bcc}}$); since the shape has volume 1/16, while $\bar{S}_1'$ has volume 1/6, we get a lattice covering of $\mathbf{R}^3$ by $\bar{S}_1'$ with efficiency 3/8.

The discrete form of this shape, scaled by a factor s_i in the ith dimension, is a subset of $\mathbf{Z}^3$ of size $4s_1s_2s_3$ which gives a lattice tiling of $\mathbf{Z}^3$; this subset is included in S_k', where $k = s_1 + s_2 + s_3 + \max(s_1, s_2, s_3) - 3$. Optimizing this for a given $k \geq 1$ gives a subset of S_k' which tiles and has size $4a_3a_4a_5$, where $a_i = \lfloor (k+i)/4 \rfloor$.

One can obtain another lattice covering of $\mathbf{R}^3$ by tetrahedra as follows. If one cuts off the four corners of a regular tetrahedron at planes passing through the midpoints of the edges (so one removes four half-size regular tetrahedra), then what is left is a regular octahedron with a volume of 1/2 that of the tetrahedron. We have a lattice giving a covering of $\mathbf{R}^3$ by this octahedron with efficiency 8/9; the same lattice therefore gives a covering of $\mathbf{R}^3$ by the original tetrahedron with efficiency 4/9.

If one uses an affine transformation to change the regular tetrahedron to the tetrahedron $\bar{S}_1'$, then the corresponding lattice will be generated by $(1/6, 1/6, 1/6)$, $(1/6, -1/2, 1/6)$, and $(1/6, 1/6, -1/2)$. One fundamental region for this lattice is an affinely distorted truncated octahedron. Another can be obtained by the method of Lemma 3.1, using an ordering $\prec$ which orders vectors primarily by the sum of their coordinates; the resulting region is shown on the right side of Figure 8.1. This shape consists of 16 cubes of edge-length 1/6 for a total volume of 2/27, which, as expected, is 4/9 of $\mathrm{vol}(\bar{S}_1') = 1/6$.

Discretizing this new shape with scale factors $s_1 \leq s_2 \leq s_3$ gives a subset of $\mathbf{Z}^3$ of size $16s_1s_2s_3$ which gives a lattice tiling of $\mathbf{Z}^3$; this subset is included in S_k', where

$k = s_1 + 2s_2 + 3s_3 - 3$. Another simple optimization shows that, for any given $k \geq 3$, we get a subset of S'_k which tiles and has size $16\hat{a}_3\hat{a}_4\hat{a}_6$, where $\hat{a}_i = \lfloor (k+i)/6 \rfloor$. For large k (in fact, for all $k \geq 30$), this new lattice gives a better covering of $\mathbf{Z}^3$ by S'_k than the preceding one did, but the preceding one sometimes does better for smaller k.

Aguiló, Fiol, and Garcia [1] also work with this shape but discretize it in a rotationally symmetric way rather than separately in each dimension; the Cayley graphs they obtain are slightly larger than the graphs of size $16\hat{a}_3\hat{a}_4\hat{a}_6$ given above but are still of the form $(2/27)k^3 + O(k^2)$.

In order to see whether these lattice coverings give close-to-optimal Cayley graphs, the authors performed a computer search for the best (smallest-diameter) directed Abelian Cayley graphs on three generators. This extends similar computations performed by Aguiló, Fiol, and Garcia [1] and by Fiduccia, Forcade, and Zito [14]. The latter paper also contains a useful upper bound: an Abelian Cayley digraph on three generators with diameter k must have size at most $3(k+3)^3/25$. This improves the obvious upper bound $|S_k|$ when $k > 7$.

Comparing the above figures with the output from the authors' computations gives a slight surprise: the best cyclic groups do substantially better than the groups from the above coverings. The data are shown in Table 8.1; here "FFZ" is the Fiduccia–Forcade–Zito upper bound, "Tor." is the size of the best toroidal mesh, "Impr." refers to the larger of the sizes obtained from the two improved constructions above, "AFG" is the size attained by the Aguiló–Fiol–Garcia construction, and the remaining columns are analogous to those of Table 7.1. The computations were run on Abelian groups of sizes up to and including 8184; this means that the entries marked with an asterisk in the n'_c column (for which the FFZ bound is greater than 8184) have not been completely proven optimal, but it is extremely likely that they are.

Note that in three cases, $k = 7, 31, 33$, the best cyclic Cayley graph was not achieved using 1 as one of the generators. If we are required to use 1 as a generator (which may be useful when actually building the corresponding loop network), then the best we can do is size 78 for $k = 7$ (with generators $1, 6, 49$), size 3178 for $k = 31$ (with generators $1, 386, 1295$), and size 3794 for $k = 33$ (with generators $1, 469, 2094$).

There is one other difference between this case and the undirected case: here there are values of k for which one can do better with general Abelian groups than with cyclic groups. The improved values obtained from noncyclic groups are shown in Table 8.2. A number of these optimal graphs are actually obtained by applying Proposition 4.2(b) to smaller Cayley graphs; for instance, the Abelian graph for $k = 17$ is obtained this way from the cyclic graph for $k = 7$, which is the reason that these two graphs give exactly the same real-covering efficiency (.504).

The values in the n'_c column of Table 8.1 are so much larger than those in the preceding two columns that it is clear that the real lattices used for the preceding columns were not optimal. This is made explicit in the last column of the table, which gives the efficiency of the real lattice covering obtained from the computed integer lattice covering via Proposition 4.1(b). For $k = 1$ and $k = 3$ these coverings are just (scaled versions of) the two coverings we explicitly constructed above, but later coverings obviously do substantially better.

The best real covering obtained from these computations is that for $k = 7$, with efficiency .504. As in the undirected case, we can reconstruct generators for the lattice from the given generating set $2, 9, 35$ for $\mathbf{Z}_{84}$; after simplification, the resulting

Table 8.1
Best directed Cayley graphs of cyclic groups, three generators.

k	$\lvert S'_k\rvert$	FFZ	Tor.	Impr.	AFG	n'_c	Generators	$n'_c/\lvert S'_k\rvert$	$n'_c/\operatorname{vol}(\bar{S}'_{k+3})$
0	1		1		1	1		1	.222222
1	4		2	4	4	4	1,2,3	1	.375000
2	10		4	4	7	9	1,3,4	.900000	.432000
3	20		8	16	16	16	1,4,5	.800000	.444444
4	35		12	16	19	27	1,4,17	.771428	.472303
5	56		18	32	31	40	1,6,15	.714286	.468750
6	84		27	32	50	57	1,13,33	.678571	.469136
7	120	120	36	48	56	84	2,9,35	.700000	.504000
8	165	159	48	72	86	111	1,31,69	.672727	.500376
9	220	207	64	128	128	138	1,11,78	.627273	.479167
10	286	263	80	128	134	176	1,17,56	.615385	.480655
11	364	329	100	144	182	217	1,13,119	.596154	.474490
12	455	405	125	192	243	273	1,14,153	.600000	.485333
13	560	491	150	256	252	340	1,90,191	.607143	.498047
14	680	589	180	288	333	395	1,35,271	.580882	.482394
15	816	699	216	432	432	462	1,29,97	.566176	.475309
16	969	823	252	432	441	560	1,215,326	.577915	.489867
17	1140	960	294	500	549	648	1,76,237	.568421	.486000
18	1330	1111	343	576	676	748	1,41,147	.562406	.484613
19	1540	1277	392	600	688	861	1,27,463	.559091	.485162
20	1771	1460	448	768	844	979	1,22,351	.552795	.482781
21	2024	1658	512	1024	1024	1140	1,45,196	.563241	.494792
22	2300	1875	576	1024	1036	1305	1,246,1030	.567391	.501120
23	2600	2109	648	1024	1228	1440	1,126,415	.553846	.491579
24	2925	2361	729	1280	1445	1616	1,56,257	.552479	.492608
25	3276	2634	810	1372	1460	1788	1,154,1452	.545788	.488703
26	3654	2926	900	1600	1715	1963	1,90,780	.537219	.482923
27	4060	3240	1000	2000	2000	2224	1,425,704	.547783	.494222
28	4495	3574	1100	2000	2015	2442	1,964,1372	.543270	.491826
29	4960	3932	1210	2048	2315	2693	1,39,942	.542944	.493103
30	5456	4312	1331	2400	2646	2920	1,540,831	.535191	.487520
31	5984	4716	1452	2400	2664	3220	7,30,2277	.538102	.491553
32	6545	5145	1584	2880	3042	3591	1,1519,2031	.548663	.502531
33	7140	5598	1728	3456	3456	3850	2,475,1177	.539216	.495113
34	7770	6078	1872	3456	3474	4191	1,748,2652	.539382	.496437
35	8436	6584	2028	3456	3906	4468	1,353,2789	.529635	.488555
36	9139	7118	2197	4032	4375	4871	1,238,1113	.532990	.492692
37	9880	7680	2366	4032	4396	5328	1,345,2344	.539271	.499500
38	10660	8270	2548	4704	4921	5698*	1,1375,2410	.534522	.496046
39	11480	8890	2744	5488	5488	6131*	1,51,1589	.534059	.496518
40	12341	9540	2940	5488	5509	6513*	1,560,5070	.527753	.491504
41	13244	10222	3150	5488	6097	6942*	1,793,1860	.524162	.488965
42	14190	10935	3375	6272	6728	7533*	1,1612,4961	.530867	.496000
43	15180	11680	3600	6272	6752	8064*	1,1377,4960	.531225	.497082

Table 8.2

Best directed Cayley graphs of Abelian groups, three generators.

k	n'_a	Group	Generators	$n'_a/\lvert S'_k\rvert$	$n'_a/\operatorname{vol}(\bar{S}'_{k+3})$
12	279	$\mathbf{Z}_{93}\times\mathbf{Z}_3$	(1,0),(9,1),(10,2)	.613187	.496000
17	672	$\mathbf{Z}_{168}\times\mathbf{Z}_2\times\mathbf{Z}_2$	(2,1,0),(9,0,0),(35,0,1)	.589474	.504000
18	752	$\mathbf{Z}_{188}\times\mathbf{Z}_4$	(1,0),(13,2),(14,1)	.565414	.487204
19	888	$\mathbf{Z}_{222}\times\mathbf{Z}_2\times\mathbf{Z}_2$	(1,0,0),(142,1,0),(180,0,1)	.576623	.500376
26	1980	$\mathbf{Z}_{330}\times\mathbf{Z}_6$	(1,0),(123,2),(234,3)	.541872	.487105
27	2268	$\mathbf{Z}_{252}\times\mathbf{Z}_3\times\mathbf{Z}_3$	(2,0,0),(9,1,0),(35,0,1)	.558621	.504000
28	2448	$\mathbf{Z}_{816}\times\mathbf{Z}_3$	(1,0),(427,0),(564,1)	.544605	.493035
29	2720	$\mathbf{Z}_{680}\times\mathbf{Z}_2\times\mathbf{Z}_2$	(1,0,0),(191,1,0),(90,0,1)	.548387	.498047
30	2997	$\mathbf{Z}_{333}\times\mathbf{Z}_3\times\mathbf{Z}_3$	(1,0,0),(31,1,0),(180,0,1)	.549304	.500376
35	4500	$\mathbf{Z}_{300}\times\mathbf{Z}_{15}$	(1,0),(3,1),(214,7)	.533428	.492054
37	5376*	$\mathbf{Z}_{336}\times\mathbf{Z}_4\times\mathbf{Z}_4$	(2,1,0),(9,0,0),(35,0,1)	.544130	.504000
41	7104*	$\mathbf{Z}_{444}\times\mathbf{Z}_4\times\mathbf{Z}_4$	(1,0,0),(364,1,0),(180,0,1)	.536394	.500376
42	7641*	$\mathbf{Z}_{2547}\times\mathbf{Z}_3$	(1,0),(256,1),(2238,2)	.538478	.503111

generating vectors are $(-2,2,2)$, $(3,-3,3)$, and $(4,3,-1)$. We now have a computer-assisted proof that the lattice generated by these vectors gives a lattice covering of $\mathbf{R}^3$ by $\bar{S}'_{10}$; but one can obtain useful extra information (as well as, perhaps, more satisfaction) by proving this directly.

Proposition 8.1. *Let L'_7 be the lattice in $\mathbf{R}^3$ generated by the vectors $(-2,2,2)$, $(3,-3,3)$, and $(4,3,-1)$; then $\bar{S}'_{10}+L'_7=\mathbf{R}^3$.*

Proof. First, note that the following vectors are in L'_7:

$$\begin{aligned}
\mathbf{v}_1 &= (-2,2,2), & \mathbf{v}_8 &= (1,-1,5) = \mathbf{v}_1+\mathbf{v}_2,\\
\mathbf{v}_2 &= (3,-3,3), & \mathbf{v}_9 &= (-1,1,7) = 2\mathbf{v}_1+\mathbf{v}_2,\\
\mathbf{v}_3 &= (4,3,-1), & \mathbf{v}_{10} &= (3,4,-6) = \mathbf{v}_3-\mathbf{v}_1-\mathbf{v}_2,\\
\mathbf{v}_4 &= (6,1,-3) = \mathbf{v}_3-\mathbf{v}_1, & \mathbf{v}_{11} &= (-7,7,1) = 2\mathbf{v}_1-\mathbf{v}_2,\\
\mathbf{v}_5 &= (5,-5,1) = \mathbf{v}_2-\mathbf{v}_1, & \mathbf{v}_{12} &= (-5,-2,8) = 2\mathbf{v}_1+\mathbf{v}_2-\mathbf{v}_3,\\
\mathbf{v}_6 &= (1,6,-4) = \mathbf{v}_3-\mathbf{v}_2, & \mathbf{v}_{13} &= (-1,8,-2) = \mathbf{v}_1-\mathbf{v}_2+\mathbf{v}_3,\\
\mathbf{v}_7 &= (2,5,1) = \mathbf{v}_1+\mathbf{v}_3, & \mathbf{v}_{14} &= (8,-1,-5) = \mathbf{v}_3-2\mathbf{v}_1.
\end{aligned}$$

For each vector $\mathbf{v}_i=(r,s,t)$, we have $1\le r+s+t\le 8$; hence, the translated tetrahedron $\bar{S}'_{10}+\mathbf{v}_i$ intersects the plane $x+y+z=10$. In fact, the intersection is a triangle whose vertices have coordinates $(10-s-t,s,t)$, $(r,10-r-t,t)$, and $(r,s,10-r-s)$.

Figure 8.2 shows the upper face of $\bar{S}'_{10}$. For each $i\le 14$, it indicates which part of this face is covered by the translate $\bar{S}'_{10}+\mathbf{v}_i$. (The face is divided into unit triangles, each of which is labeled by the value(s) of i for which $\bar{S}'_{10}+\mathbf{v}_i$ covers that triangle.) Clearly, each unit triangle is labeled, so the translates $\bar{S}'_{10}+\mathbf{v}_i$ cover the entire upper face of $\bar{S}'_{10}$.

In fact, since each $\mathbf{v}_i$ has coordinates summing to at least 1, the smaller translates $\bar{S}'_9+\mathbf{v}_i$, $i\le 14$, cover the upper face of $\bar{S}'_{10}$. For any $\mathbf{x}$ in this upper face, there is an i such that $\mathbf{x}\in\bar{S}'_9+\mathbf{v}_i$, so $\bar{S}'_1+\mathbf{x}\subseteq\bar{S}'_1+\bar{S}'_9+\mathbf{v}_i=\bar{S}'_{10}+\mathbf{v}_i$. Since $\bar{S}'_{11}$ is the union of $\bar{S}'_{10}$ and the sets $\bar{S}'_1+\mathbf{x}$ for $\mathbf{x}$ in the upper face of $\bar{S}'_{10}$, we have $\bar{S}'_{11}\subseteq\bar{S}'_{10}+L'_7$, and

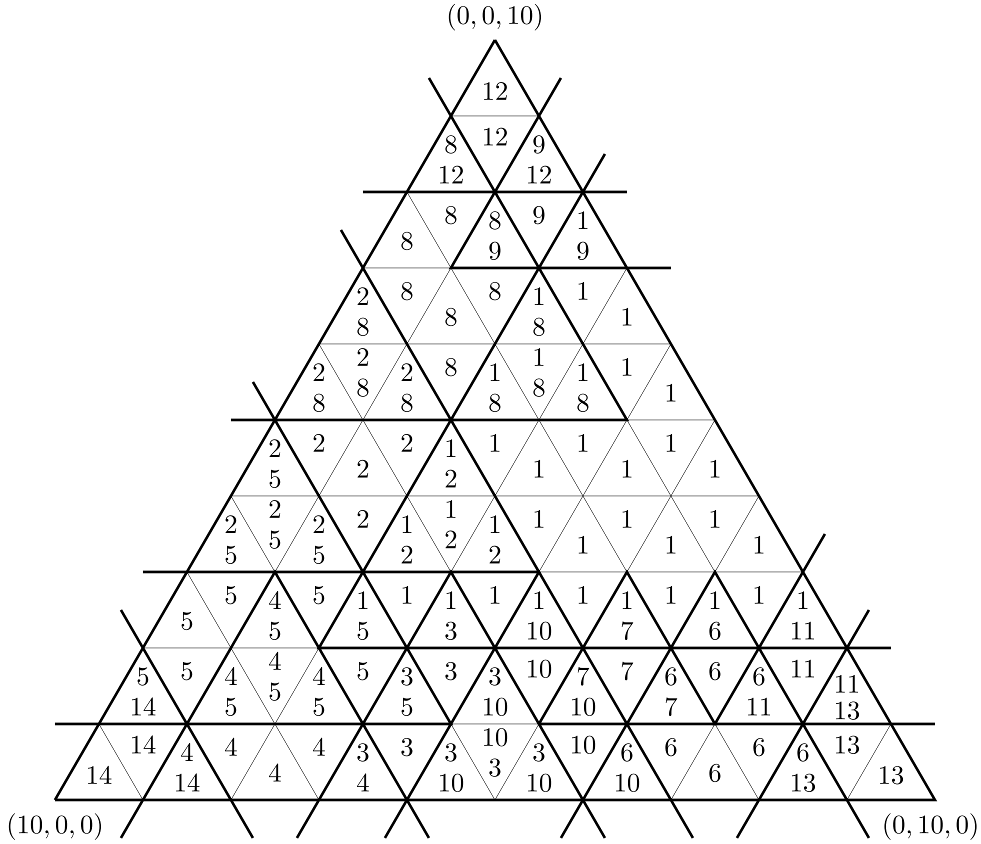


FIG. 8.2. *Coverage of one face of $\bar{S}'_{10}$ under L'_7.*

hence $\bar{S}'_{11} + L'_7 \subseteq \bar{S}'_{10} + L'_7$.

We now prove by induction that, for all integers $k \geq 10$, $\bar{S}'_k + L'_7 \subseteq \bar{S}'_{10} + L'_7$. The case $k = 10$ is trivial. If it is true for k, then

$$\bar{S}'_{k+1} + L'_7 = \bar{S}'_1 + \bar{S}'_k + L'_7 \subseteq \bar{S}'_1 + \bar{S}'_{10} + L'_7 = \bar{S}'_{11} + L'_7 \subseteq \bar{S}'_{10} + L'_7,$$

so it is true for $k + 1$.

Finally, for any $\mathbf{y} \in \mathbf{R}^3$, there is a member $\mathbf{w}$ of L'_7 such that the coordinates of $\mathbf{y} - \mathbf{w}$ are all positive (e.g., let $\mathbf{w}$ be a large multiple of $-\mathbf{v}_7$). Then $\mathbf{y} - \mathbf{w} \in \bar{S}'_k$ for some k, so $\mathbf{y} \in \bar{S}'_k + L'_7$, and hence $\mathbf{y} \in \bar{S}'_{10} + L'_7$. Therefore, $\bar{S}'_{10} + L'_7 = \mathbf{R}^3$. □

This covering has efficiency $\det(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)/\operatorname{vol}(\bar{S}'_{10}) = .504$. Hence, Corollary 4.7(b) gives the following.

COROLLARY 8.2. *For all k, there is a directed Cayley graph of an Abelian group on three generators which has diameter k and size at least $0.084k^3 + O(k^2)$.*

We can now use the method of (the real version of) Lemma 3.1 to get a fundamental region $T'_7 \subseteq \bar{S}'_{10}$ for the lattice L'_7. (Recall that any such region must have volume $\det(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = 84$.) To do this, just start with $\bar{S}'_{10}$, look at each of the vectors $\mathbf{v}_i$ ($i \leq 14$) defined above, and delete those points of $\bar{S}'_{10}$ which lie in $\bar{S}'_{10} + \mathbf{v}_i$.

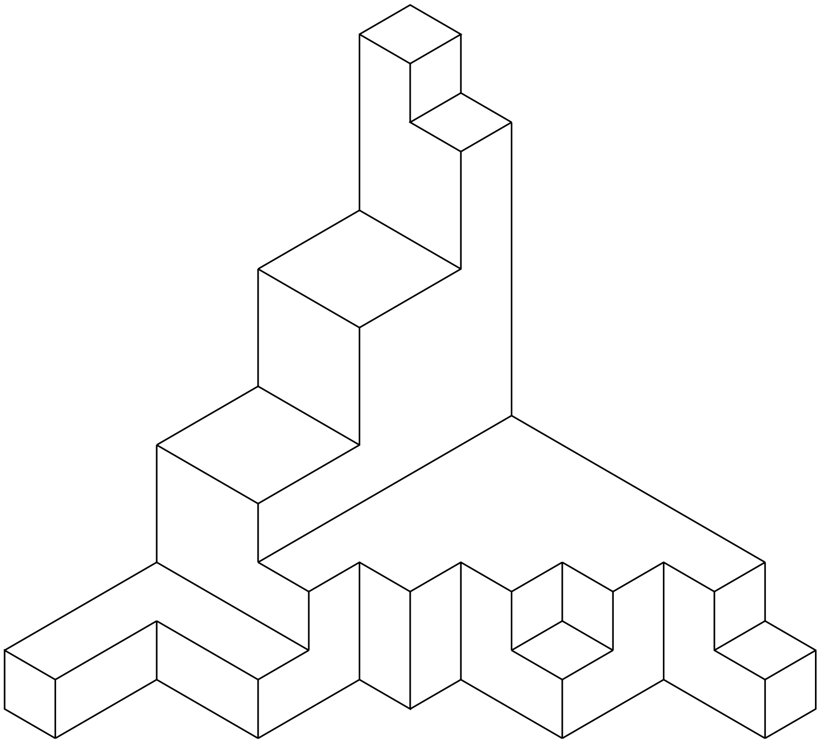


FIG. 8.3. *A subset of the tetrahedron $\bar{S}'_{10}$ which tiles space.*

(We have $\mathbf{v}_i \succ \mathbf{0}$ for all i if $\prec$ orders vectors primarily by the sum of coefficients.) What is left is the set shown in Figure 8.3; since this is the union of 84 unit cubes, we know that there is no need to subtract further translates $\bar{S}'_{10} + \mathbf{w}$. This set was obtained independently by Fiduccia, Forcade, and Zito [14].

So this set T'_7 gives a lattice tiling of $\mathbf{R}^3$ using L'_7. This tiling is quite unusual; the translates of T'_7 fit together in a peculiar way, seeming to wind around each other. One interesting fact is that each translate of T'_7 is adjacent to (i.e., shares a boundary segment of positive area with) 28 other translates, a surprisingly high number. (T'_7 itself is adjacent to $T'_7 + \mathbf{v}_i$ and $T'_7 - \mathbf{v}_i$ for $i \leq 14$.)

In many of the tilings we constructed explicitly, there was a polycube fundamental region like T'_7, but there was also an alternate fundamental region which was convex; for instance, for the optimal covering of $\mathbf{R}^2$ by right triangles, one could use either an L-tromino or a hexagon as the fundamental region. Clearly L'_7 has convex fundamental regions (e.g., its Voronoi regions), but it turns out that they are unsuitable for the current problem, as shown below.

PROPOSITION 8.3. *There is no convex fundamental region for the lattice L'_7 included within the tetrahedron $\bar{S}'_{10}$.*

Proof. Since T'_7 gives a lattice tiling of $\mathbf{R}^3$ by L'_7, every point of $\mathbf{R}^3$, except for those lying on boundaries of the tiling, can be translated by a vector in L'_7 to a unique point of T'_7. In particular, if we look at the part of $\bar{S}'_{10}$ lying outside T'_7, then we can break it up into finitely many parts (in fact, we can just cut it along the integer translates of the three coordinate planes) which can be translated in a unique way by members of L'_7 so as to lie within T'_7.

If one does this, one finds that there are parts of T'_7 which do not get covered by translates of parts of $\bar{S}'_{10} \setminus T'_7$. Most of these uncovered parts look like inverted copies of $\bar{S}'_1$ (i.e., translates of $-\bar{S}'_1$), although there are some larger ones. In particular, the sets $(1,1,1)-\bar{S}'_1$, $(8,1,1)-\bar{S}'_1$, $(1,8,1)-\bar{S}'_1$, and $(1,1,8)-\bar{S}'_1$ are not covered by such translates. This implies that each of those four sets is disjoint (except for boundaries) from all of the translates $\bar{S}'_{10}+\mathbf{w}$ for $\mathbf{w} \in L'_7 \setminus \{\mathbf{0}\}$. It follows that any fundamental region for L'_7 included within $\bar{S}'_{10}$ must include all four of these sets.

If the fundamental region is also convex, then it must contain any convex combinations of points in those four sets; in particular, it must include the sets $(3,1,1)-\bar{S}'_1$ and $(1,3,3)-\bar{S}'_1$. But $(1,3,3)=(3,1,1)+\mathbf{v}_1$, so the $\mathbf{v}_1$-translate of the region overlaps the region itself in a set of positive volume, which is impossible for a fundamental region of L'_7. Therefore, no fundamental region of L'_7 within $\bar{S}'_{10}$ can be convex. □

There is no obvious reason why the lattice L'_7 should be exactly optimal for a lattice covering of $\mathbf{R}^3$ by the tetrahedron $\bar{S}'_{10}$. In the case of undirected graphs on three generators, the lattices obtained for each k were not optimal but were closer and closer approximations to the lattice L_{bcc}, which does appear to be optimal. One would expect something similar to occur in the directed case, but it does not; the real lattice efficiencies in the last columns of Tables 8.1 and 8.2 go up and down irregularly and (so far) do not exceed the value of .504 attained by L'_7.

Given this, it seems reasonable to examine L'_7 and try to adjust it slightly in order to improve its efficiency; there should be some locally optimal lattice to which L'_7 is an approximation, and we would like to find it. Quite surprisingly, it turns out that no adjustment is necessary. Just before submitting the present paper, the authors found a recent paper of Forcade and Lamoreaux [15] proving this same result by a method slightly different from that presented here.

THEOREM 8.4 (Forcade and Lamoreaux [15]). *Among those lattices L for which $\bar{S}'_{10}+L=\mathbf{R}^3$, the lattice L'_7 is locally optimal.*

Proof. We use the same methods as for Theorem 7.2. Recall the vectors $\mathbf{v}_1,\dots,\mathbf{v}_{14}$ from Proposition 8.1. The vectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ generate L'_7; a nearby lattice L will be generated by nearby vectors $\mathbf{v}'_1=(a_1,b_1,c_1)$, $\mathbf{v}'_2=(a_2,b_2,c_2)$, and $\mathbf{v}'_3=(a_3,b_3,c_3)$. Again concatenate $\mathbf{v}'_1$, $\mathbf{v}'_2$, $\mathbf{v}'_3$ and $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ to get $\mathbf{v}'$ and $\mathbf{v}$ in $\mathbf{R}^9$. Let $F(\mathbf{v}')$ be the determinant of the matrix with rows $\mathbf{v}'_1$, $\mathbf{v}'_2$, $\mathbf{v}'_3$; then we have $F(\mathbf{v})=84$, and we want to see that $F(\mathbf{v}')<84$ for any other $\mathbf{v}'$ near $\mathbf{v}$ for which the corresponding lattice L satisfies $\bar{S}'_{10}+L=\mathbf{R}^3$. We compute that the gradient of F at the point $\mathbf{v}$ is $\mathbf{g}=(-6,15,21,8,-6,14,12,12,0)$.

Referring back to Figure 8.2, we see that the point $(1,1,8)$ is on the boundary of $\bar{S}'_{10}+\mathbf{v}_i$ for $i=8,9,12$, as well as on the boundary of $\bar{S}'_{10}$ itself; one can check that no other L'_7-translate of $\bar{S}'_{10}$ is near this point. The nearby lattice L contains points $\mathbf{0}$, $\mathbf{v}'_8=\mathbf{v}'_1+\mathbf{v}'_2$, $\mathbf{v}'_9=2\mathbf{v}'_1+\mathbf{v}'_2$, and $\mathbf{v}'_{12}=2\mathbf{v}'_1+\mathbf{v}'_2-\mathbf{v}'_3$. For any small positive ε, the point $(a_1+a_2-\varepsilon, 2b_1+b_2-\varepsilon, 2c_1+c_2-c_3-\varepsilon)$, which is near $(1,1,8)$, will not be in $\mathbf{v}'_8+\bar{S}'_{10}$ because its first coordinate is smaller than that of $\mathbf{v}'_8$. By looking at second and third coordinates respectively, we see that this point is not in $\mathbf{v}'_9+\bar{S}'_{10}$ or $\mathbf{v}'_{12}+\bar{S}'_{10}$ either. Hence, in order to have $\bar{S}'_{10}+L=\mathbf{R}^3$, this point must be in $\bar{S}'_{10}$ itself, so we must have

$$a_1+a_2+2b_1+b_2+2c_1+c_2-c_3-3\varepsilon \le 10.$$

Since ε can be arbitrarily small, we need

$$a_1+a_2+2b_1+b_2+2c_1+c_2-c_3 \le 10$$

in order to have $\bar{S}'_{10} + L = \mathbf{R}^3$. So we have the constraint

$$\mathbf{u}_1 \cdot \mathbf{v}' \leq 10, \quad \text{where } \mathbf{u}_1 = (1,2,2,1,1,1,0,0,-1).$$

The same reasoning applied at the points $(1,2,7)$, $(3,2,5)$, $(5,2,3)$, $(5,3,2)$, $(4,4,2)$, $(3,5,2)$, $(2,6,2)$, $(1,7,2)$, $(8,1,1)$, $(6,3,1)$, $(3,6,1)$, and $(1,8,1)$ gives the constraints

$$\begin{aligned}
\mathbf{u}_2 \cdot \mathbf{v}' &\leq 10, &&\text{where } \mathbf{u}_2 = (1,1,2,1,0,1,0,0,0),\\
\mathbf{u}_3 \cdot \mathbf{v}' &\leq 10, &&\text{where } \mathbf{u}_3 = (0,1,1,1,0,1,0,0,0),\\
\mathbf{u}_4 \cdot \mathbf{v}' &\leq 10, &&\text{where } \mathbf{u}_4 = (-1,1,0,1,0,1,0,0,0),\\
\mathbf{u}_5 \cdot \mathbf{v}' &\leq 10, &&\text{where } \mathbf{u}_5 = (-1,0,1,1,0,0,0,1,0),\\
\mathbf{u}_6 \cdot \mathbf{v}' &\leq 10, &&\text{where } \mathbf{u}_6 = (0,-1,1,0,-1,0,1,1,0),\\
\mathbf{u}_7 \cdot \mathbf{v}' &\leq 10, &&\text{where } \mathbf{u}_7 = (-1,1,1,-1,0,0,1,1,0),\\
\mathbf{u}_8 \cdot \mathbf{v}' &\leq 10, &&\text{where } \mathbf{u}_8 = (1,0,1,0,-1,0,1,1,0),\\
\mathbf{u}_9 \cdot \mathbf{v}' &\leq 10, &&\text{where } \mathbf{u}_9 = (0,2,1,-1,-1,0,1,0,0),\\
\mathbf{u}_{10} \cdot \mathbf{v}' &\leq 10, &&\text{where } \mathbf{u}_{10} = (-2,-1,-1,0,0,1,1,1,0),\\
\mathbf{u}_{11} \cdot \mathbf{v}' &\leq 10, &&\text{where } \mathbf{u}_{11} = (-1,0,-1,0,0,1,1,1,0),\\
\mathbf{u}_{12} \cdot \mathbf{v}' &\leq 10, &&\text{where } \mathbf{u}_{12} = (-1,0,1,-1,-1,0,1,1,1),\\
\mathbf{u}_{13} \cdot \mathbf{v}' &\leq 10, &&\text{where } \mathbf{u}_{13} = (0,1,2,-1,-1,-1,1,1,0).
\end{aligned}$$

Again note that all thirteen of these inequalities are satisfied with equality when $\mathbf{v}' = \mathbf{v}$. Hence, they can be rewritten as $\mathbf{u}_i \cdot (\mathbf{v}' - \mathbf{v}) \leq 0$ for $i = 1, 2, \ldots, 13$.

One can easily check that the vectors $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_4$, $\mathbf{u}_5$, $\mathbf{u}_6$, $\mathbf{u}_7$, $\mathbf{u}_8$, $\mathbf{u}_{10}$ are linearly independent; their common null space is generated by the vector $\mathbf{w} = (0,0,0,1,1,-1,2,-1,1)$. (The other five vectors $\mathbf{u}_i$ are also orthogonal to $\mathbf{w}$, so they are linear combinations of the eight listed above.) Also, we have

$$\mathbf{g} = \mathbf{u}_1 + 4.8\mathbf{u}_2 + 6.4\mathbf{u}_4 + \mathbf{u}_5 + 1.6\mathbf{u}_6 + 3.2\mathbf{u}_7 + 3.4\mathbf{u}_8 + \mathbf{u}_9 + 1.8\mathbf{u}_{10} + \mathbf{u}_{12}.$$

Let C be the closed cone consisting of all vectors $\mathbf{t}$ in the subspace spanned by $\mathbf{u}_1, \ldots, \mathbf{u}_{13}$ such that $\mathbf{u}_i \cdot \mathbf{t} \leq 0$ for all $i \leq 13$. Then the above equations imply that $\mathbf{g} \cdot \mathbf{t} \leq 0$ for all $\mathbf{t}$ in C, and equality can hold only when $\mathbf{t} = \mathbf{0}$. Hence, as in Theorem 7.2, there is a neighborhood U of $\mathbf{g}$ such that, for any $\mathbf{g}'$ in U and any nonzero $\mathbf{t}$ in C, $\mathbf{g}' \cdot \mathbf{t} < 0$.

We can compute that, for any real number r, the determinant for the lattice given by $\mathbf{v} + r\mathbf{w}$ is $F(\mathbf{v} + r\mathbf{w}) = 84 - 12r^2$. Clearly, this is at most 84, with equality only when $r = 0$.

Now, any vector $\mathbf{v}'$ close to $\mathbf{v}$ can be expressed as $\mathbf{v} + \mathbf{t}_1 + \mathbf{t}_2$ where $\mathbf{t}_1$ is a small multiple of $\mathbf{w}$ and $\mathbf{t}_2$ is a small linear combination of the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_{13}$. The reasoning from Theorem 7.2 shows that $\mathbf{t}_2$ must be in C if $\bar{S}'_{10} + L = \mathbf{R}^3$. Also as in that theorem, we find that $F(\mathbf{v} + \mathbf{t}_1) \leq F(\mathbf{v})$ with equality only when $\mathbf{t}_1 = \mathbf{0}$, and $F(\mathbf{v} + \mathbf{t}_1 + \mathbf{t}_2) \leq F(\mathbf{v} + \mathbf{t}_1)$ with equality only when $\mathbf{t}_2 = \mathbf{0}$. Therefore, $F(\mathbf{v}') \leq F(\mathbf{v})$, with equality holding only when $\mathbf{v}' = \mathbf{v}$. So $\mathbf{v}$ gives a local maximum of F, as desired. □

This and the computational evidence make it plausible that L'_7 actually gives an optimal lattice covering of $\mathbf{R}^3$ by $\bar{S}'_{10}$, and hence that the asymptotic formula in Corollary 8.2 is optimal.

9. Cayley graphs on more than three generators. In higher dimensions, analogues of many of the preceding constructions exist, but they do not produce lattice coverings as efficient as one would hope for.

For lattice coverings with the d-dimensional dual cube, one can use the d-dimensional body-centered cubic lattice (the set of vectors in $\mathbf{Z}^d$ whose coordinates are all odd or all even). By the same argument as for the three-dimensional case, this lattice gives a lattice covering of $\mathbf{R}^d$ by $\bar{S}_{d/2}$. The efficiency of this covering is $2^{d-1}/\operatorname{vol}(\bar{S}_{d/2}) = 2^{d-1}d!/d^d$, which is 2^{d-1} times the efficiency of the covering using the ordinary cubic lattice $\mathbf{Z}^d$.

As usual, the Cayley graph corresponding to this lattice is a twisted toroidal mesh. For a given number m, one can connect the elements of $\mathbf{Z}_{2m}^{d-1} \times \mathbf{Z}_m$ as in an ordinary toroidal mesh, except that the wraparound connections for the last coordinate are twisted along all of the other coordinates: $(x_1, \ldots, x_{d-1}, m-1)$ is connected to $(x_1 \pm m, \ldots, x_{d-1} \pm m, 0)$. This gives a graph of diameter $\lfloor dm/2 \rfloor$ and size $2^{d-1}m^d$, which is about 2^{d-1} times as large as the best ordinary toroidal mesh of this diameter.

One can optimize this slightly. Given the dimension d and the desired diameter k, let q and r be the quotient and remainder when $2k+1$ is divided by d; we assume k is large enough that $q > 0$. Then a good lattice L to use is the body-centered cubic lattice above, scaled up by a factor $q+1$ in each of the first r coordinates and by a factor q in the remaining $d-r$ coordinates. The resulting $\mathbf{Z}^d/L$ is isomorphic to $(\mathbf{Z}_{2q+2}^r \times \mathbf{Z}_{2q}^{d-r})/H$ with the canonical generators, where H is the two-element subgroup $\{\mathbf{0}, (q+1, \ldots, q+1, q, \ldots, q)\}$ (there are r $q+1$'s); it can be laid out as a twisted toroidal mesh on $\mathbf{Z}_{2q+2}^r \times \mathbf{Z}_{2q}^{d-r-1} \times \mathbf{Z}_q$ or on $\mathbf{Z}_{2q+2}^{r-1} \times \mathbf{Z}_{2q}^{d-r} \times \mathbf{Z}_{q+1}$. If q is even and $r > 0$, this Cayley graph is isomorphic to that of $\mathbf{Z}_{q+1} \times \mathbf{Z}_{2q+2}^{r-1} \times \mathbf{Z}_{2q}^{d-r}$ with the generators $\mathbf{e}_2, \ldots, \mathbf{e}_d$ and $(1, \ldots, 1, q, \ldots, q)$ with r 1's; if q is odd or $r = 0$, then it is isomorphic to the Cayley graph of $\mathbf{Z}_{2q+2}^r \times \mathbf{Z}_{2q}^{d-r-1} \times \mathbf{Z}_q$ with generators $\mathbf{e}_1, \ldots, \mathbf{e}_{d-1}$ and $(q+1, \ldots, q+1, 1, \ldots, 1)$ with r $q+1$'s. The size of this graph is slightly larger than the size of the cyclic Cayley graph constructed by Chen and Jia [5], but the ratio of the two sizes tends to 1 for large k.

For the directed case, we must consider lattice coverings by d-simplices; as usual, by affine invariance, it doesn't matter which simplex is used. One can show that a lattice for covering with a given d-simplex is given by the following generating vectors: for each face of the simplex, take a vector which is twice the vector from the centroid of the simplex to the centroid of that face. (This gives $d+1$ vectors, but they sum to $\mathbf{0}$, so just take d of them.) The efficiency of this covering works out to be $d!2^d/(d^d(d+1))$.

Unfortunately, in both cases, the efficiency decreases exponentially with d: by Stirling's formula,

$$\frac{2^{d-1}d!}{d^d} \sim \sqrt{\frac{\pi d}{2}} \left(\frac{2}{e}\right)^d$$

and

$$\frac{d!2^d}{d^d(d+1)} \sim \sqrt{\frac{2\pi}{d}} \left(\frac{2}{e}\right)^d.$$

This seems to be the case for all known explicitly constructed lattice coverings by these shapes (and by spheres).

On the other hand, in 1959 Rogers [19] gave a nonconstructive proof that there exist much more efficient lattice coverings by these shapes (or by any convex body) in high dimensions, and he gave an even better result for the case of spheres. More recently, Gritzmann [16] extended the latter result to apply to any convex body with a sufficient number of mutually orthogonal hyperplanes of symmetry. (The number required is quite small: only $\lfloor \log_2 \ln d \rfloor + 5$.) Gritzmann's result states that there is a constant c (not depending on d or on the convex body) such that, for any convex body K in $\mathbf{R}^d$ with the above number of mutually orthogonal planes of symmetry, there is a lattice covering of $\mathbf{R}^d$ by K with density at most $cd(\ln d)^{1+\log_2 e}$.

The regular dual d-cube and the regular d-simplex do have the required symmetry planes for large enough d. This is clear for the dual d-cube; it has the same d orthogonal planes of symmetry as the d-cube to which it is dual. For the regular d-simplex, note that the perpendicular bisector of an edge is a hyperplane of symmetry and that edges which do not share a vertex have orthogonal directions (the easiest way to see this is to look at the regular d-simplex in $\mathbf{R}^{d+1}$ whose vertices are $\mathbf{e}_1, \dots, \mathbf{e}_{d+1}$, and take dot products), so one can find $\lceil d/2 \rceil$ mutually orthogonal hyperplanes of symmetry. Therefore, we get lattice coverings of the specified density for large d, and by adjusting the constant c we can make the bound apply for all d (for these two particular shapes). Therefore, letting $\bar{c} = c^{-1}$, we can use Corollary 4.7 to get the following.

Theorem 9.1. *There is a constant $\bar{c} > 0$ (not depending on d or k) such that, for any fixed $d > 1$ and for all k, there exist undirected Cayley graphs of Abelian groups on d generators having diameter $\leq k$ and size at least*

$$\frac{2^d \bar{c}}{d! d (\ln d)^{1+\log_2 e}} k^d + O(k^{d-1}),$$

and there exist directed Cayley graphs of Abelian groups on d generators having diameter $\leq k$ and size at least

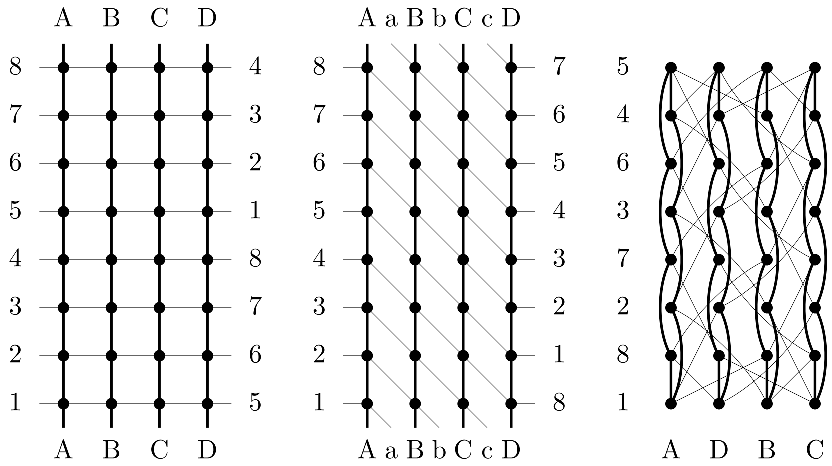
$$\frac{\bar{c}}{d! d (\ln d)^{1+\log_2 e}} k^d + O(k^{d-1}).$$

The coverings produced by this method are probably fairly strange. We seem to have already run into this in three dimensions for the directed case; for the undirected case it apparently happens later.

10. Layouts with short wires. The obvious way to lay out a toroidal mesh is as a rectangular array with mesh connections between adjacent nodes in the array and with long wires connecting opposite ends of the array; these long wires may cause communication delays. However, there is a standard trick for rearranging the layout so as to remove the need for long wires. In the one-dimensional case, instead of placing the nodes in the order $1, 2, 3, \dots, n$ (where i is connected to $i+1$ for $i = 1, 2, \dots, n-1$ and n is connected to 1), one can place them in the order $1, n, 2, n-1, 3, n-2, \dots$; then the maximum required wire length is only twice the mesh spacing. In higher dimensions, one can apply the same trick to each dimension separately, and again the required wire length is twice the mesh spacing.

It is not immediately obvious that this interleaving trick can be applied to twisted toroidal meshes; a simple interleaving in each dimension would not make the twisted cross-connections short. But it is possible to get short-wire layouts for the twisted meshes in a similar way. One approach is to perform the interleaving *twice* on the long dimensions of the mesh; for instance, if the mesh has length 16 in one of the long dimensions, then the nodes would be arranged in the order

$$1, 9, 16, 8, 2, 10, 15, 7, 3, 11, 14, 6, 4, 12, 13, 5.$$



(a) (b) (c)

FIG. 10.1. *Short-wire layout for a twisted toroidal mesh.*

Then wires in this dimension would have length at most four times the mesh spacing. Now, when one does a single interleaving on the short dimension, the twisted cross-connections become short as well.

Another method is shown in Figure 10.1(a)–(c). Here the idea is to modify the original arrangement (a) by shearing the mesh (rotating the ith level in the short dimension by $i-1$ units in each of the long dimensions), as shown in (b), so that the twisted cross-connections become (almost) straight. Then one can do an ordinary interleaving in each dimension to get the result shown in (c). This gives a layout in which the maximum wire length is $2\sqrt{d}$ times the mesh spacing.

Some of the other Abelian Cayley graphs we have considered can be treated similarly, especially the ones which differ only slightly from twisted toroidal meshes. For the optimal two-generator undirected Abelian Cayley graph, if one starts with the almost-rectangular layout shown in Figure 5.1 (a $(k{+}1) \times (k{+}1)$ square next to a $k \times k$ square) and performs a shear as in Figure 10.1, then the result is a $2k \times (k{+}1)$ rectangle with one node left over; this can then be interleaved to get a short-wire layout. A more difficult case is the two-generator directed graph from Theorem 6.2 and Corollary 6.3; here one can start with the natural L-shaped layout and perform shears on separate parts to obtain a rectangle with dimensions $(a+2b) \times a$ (made up of three subrectangles with different shear patterns), where the necessary cross-connections are almost straight across, and hence interleaving will give a good layout.

11. Generators of order 2. The undirected Cayley graphs produced so far all have even degree (twice the number of generators). If one is interested in undirected Cayley graphs of odd degree, one will have to use a generator of order 2.

Using d unrestricted generators plus one order-2 generator, one can get an undirected Abelian Cayley graph of a given diameter which is about twice as large as one can get using d unrestricted generators alone. More precisely, if $n_a(d,k)$ is the size of the largest Abelian Cayley graph of diameter k using d generators, and $n_a^+(d,k)$ is the size of the largest such graph using d generators plus one order-2 generator, then

$$2n_a(d,k-1) \leq n_a^+(d,k) \leq 2n_a(d,k).$$

To see this, first let G be generated by $g_1, \dots, g_d$ and ρ, where ρ has order 2. If G has diameter at most k using these generators, and H is the subgroup of size 2 generated by ρ, then G/H is generated by the images $g_i + H$ for $1 \le i \le d$ with diameter at most k, so $|G/H| \le n_a(d,k)$, so $|G| \le 2n_a(d,k)$; hence, $n_a^+(d,k) \le 2n_a(d,k)$. On the other hand, if G is generated by $g_1, \dots, g_d$ with diameter at most $k-1$, then $G \times \mathbf{Z}_2$ is generated by $(g_i, 0)$ for $i = 1, \dots, d$ and $(0,1)$ and has diameter at most k using these generators; this shows that $2n_a(d,k-1) \le n_a^+(d,k)$.

We can also study $n_a^+(d,k)$ using the same methods that were used for $n_a(d,k)$. The appropriate free (universal) group to use here is $\mathbf{Z}^d \times \mathbf{Z}_2$, with the canonical generators $(\mathbf{e}_i, 0)$ for $i = 1, \dots, d$ and $(\mathbf{0}, 1)$. The set of elements of this group which can be written as a word of length at most k in the generators is precisely $W_k = (S_k \times 0) \cup (S_{k-1} \times 1)$. For any Abelian group G with generators $g_1, \dots, g_d$ and ρ (ρ of order 2), there is a unique homomorphism from $\mathbf{Z}^d \times \mathbf{Z}_2$ to G taking $(\mathbf{e}_i, 0)$ to g_i and $(\mathbf{0}, 1)$ to ρ; the Cayley graph of G using these generators has diameter at most k if and only if the homomorphism maps W_k onto G. So the obvious upper limit for the size of G is $|W_k| = |S_k| + |S_{k-1}|$.

We are now led to study quotient groups $(\mathbf{Z}^d \times \mathbf{Z}_2)/N$, where N is a (normal) subgroup of $\mathbf{Z}^d \times \mathbf{Z}_2$ of finite index; we want such an N of the largest possible index such that $W_k + N = \mathbf{Z}^d \times \mathbf{Z}_2$.

One simple possibility is that $N \subseteq \mathbf{Z}^d \times \{0\}$; in this case the resulting group is just $(\mathbf{Z}^d/N_0) \times \mathbf{Z}_2$, where $N = N_0 \times \{0\}$. It is easy to see that the diameter of this group is precisely one more than the diameter of $\mathbf{Z}^d/N_0$ using the canonical d generators.

Note that N_0 is a d-dimensional lattice; let $\mathbf{v}_1, \dots, \mathbf{v}_d$ be a list of generators for this lattice. Now let N' be the subgroup of $\mathbf{Z}^d \times \mathbf{Z}_2$ generated by $(\mathbf{v}_i, 1)$ for $i = 1, \dots, d$. Then we have

$$|\mathbf{Z}^d \times \mathbf{Z}_2 : N'| = 2\,|\mathbf{Z}^d : N_0| = |\mathbf{Z}^d \times \mathbf{Z}_2 : N|\,.$$

Furthermore, the diameter of $(\mathbf{Z}^d \times \mathbf{Z}_2)/N'$ is at most one more than the diameter of $\mathbf{Z}^d/N_0$, which means that it is no larger than the diameter of $(\mathbf{Z}^d \times \mathbf{Z}_2)/N$.

This shows that, when trying to determine $n_a^+(d,k)$, we may restrict ourselves to studying subgroups N of $\mathbf{Z}^d \times \mathbf{Z}_2$ of finite index which are *not* included in $\mathbf{Z}^d \times \{0\}$.

So choose $\mathbf{g} \in \mathbf{Z}^d$ such that $(\mathbf{g}, 1) \in N$. The subgroup $N \cap (\mathbf{Z}^d \times \{0\})$ is of index 2 in N and hence of finite index in $\mathbf{Z}^d \times \mathbf{Z}_2$. So we have $N \cap (\mathbf{Z}^d \times \{0\}) = L \times \{0\}$ for some d-dimensional lattice L. Note that $(2\mathbf{g}, 0) = 2(\mathbf{g}, 1) \in N$, so $2\mathbf{g} \in L$. (Normally $\mathbf{g}$ will not be in L; if $\mathbf{g} \in L$, then $(\mathbf{g}, 0) \in N$, so $(\mathbf{0}, 1) \in N$, and so the order-2 generator collapses to the identity in the quotient group.) Also, we have $|\mathbf{Z}^d \times \mathbf{Z}_2 : N| = |\mathbf{Z}^d : L|$.

It is now easy to see that

$$(W_k + N) \cap (\mathbf{Z}^d \times \{0\}) = ((S_k + L) \cup (S_{k-1} + \mathbf{g} + L)) \times \{0\}.$$

Hence, in order to have $W_k + N = \mathbf{Z}^d \times \mathbf{Z}_2$, it is necessary to have

$$(S_k + L) \cup (S_{k-1} + \mathbf{g} + L) = \mathbf{Z}^d. \qquad (*)$$

This necessary condition is also sufficient because

$$\begin{aligned}(W_k + N) \cap (\mathbf{Z}^d \times \{1\}) &= ((S_{k-1} + L) \cup (S_k + \mathbf{g} + L)) \times \{1\} \\ &= (((S_k + L) \cup (S_{k-1} + \mathbf{g} + L)) + \mathbf{g}) \times \{1\}.\end{aligned}$$

So our goal is to find such a lattice L and extra generator $\mathbf{g}$ (with $2\mathbf{g} \in L$) so that $(*)$ is satisfied and $|\mathbf{Z}^d : L|$ is as large as possible.

We are now ready to consider specific values of d. As usual, the case $d = 1$ is easy. The maximal possible value of $|\mathbf{Z} : L|$ is $|W_k| = 4k$, and this value is attained by letting L be generated by the element $4k$, with $\mathbf{g} = 2k$. This leads to the cyclic Cayley graph on the group $\mathbf{Z}_{4k}$ with unrestricted generator 1 and order-2 generator $2k$.

For $d = 2$ we have a situation very similar to that in Figure 5.1 (lattice coverings with Aztec diamonds), but not identical because we must use two different shapes. The upper bound on $|\mathbf{Z}^2 : L|$ is $|W_k| = 4k^2 + 2$. For $k = 1$ this bound is actually attainable; it leads to the Cayley graph from $\mathbf{Z}_6$ with unrestricted generators 1 and 2 and order-2 generator 3. But for $k > 1$ the pieces S_k and S_{k-1} do not fit together well enough to give a perfect tiling. The best one can do is the lattice L generated by $(2k + 1, 1)$ and $(-1, 2k - 1)$, with the extra generator $\mathbf{g} = (k, k)$, as shown in Figure 11.1.

The graph of diameter k resulting from this covering is the Cayley graph of the cyclic group $\mathbf{Z}_{4k^2}$ with unrestricted generators 1 and $2k-1$ and order-2 generator $2k^2$. One can get another Cayley graph of this size by using the lattice generated by $(2k, 0)$ and $(0, 2k)$, but this graph will not be cyclic; it comes from the group $\mathbf{Z}_{2k} \times \mathbf{Z}_{2k}$.

The outlined shape in Figure 11.1 (a $(2k + 1) \times (2k - 1)$ rectangle with one extra point) is a fundamental region which is convenient for an actual layout of nodes in a network. Note that a $2k \times 2k$ rectangle (or, for that matter, any rectangle with both sides greater than 1) will not work as a fundamental region for this lattice. The alternate lattice in the previous paragraph does allow one to use a layout which is a $2k \times 2k$ rectangle; in fact, this is just a toroidal mesh. However, in either case one will have to make the extra connections specified by the order-2 generator.

When one moves to $d = 3$, it becomes harder to get optimal results, so again the authors resorted to a computational search. For $k = 1$, the best graph is the Cayley graph of $\mathbf{Z}_8$ with unrestricted generators $1, 2, 3$ and order-2 generator 4; for $k = 2$, the best is $\mathbf{Z}_{26}$ with unrestricted generators $1, 2, 8$ and order-2 generator 13. For $3 \leq k \leq 10$, the optimal results, like those for three generators alone, form a pattern of period 3, as shown in Table 11.1. (Again the best graphs are all cyclic. This time the parameter a is defined to be the integer nearest $2k/3$.)

We can now apply the methods in the proof of Theorem 7.1 to show the following.

THEOREM 11.1. *For each $k \geq 3$, the cyclic undirected Cayley graph using the group and generators specified in Table* 11.1 *has diameter k.*

The authors again conjecture that these are actually optimal such Abelian Cayley graphs for all $k \geq 3$, not just for $3 \leq k \leq 10$.

For $d > 3$, one can get reasonably good results by letting L be approximately a cubic lattice, with $\mathbf{g}$ near the center of one of the cubes; this makes $L \cup (\mathbf{g} + L)$ approximately a body-centered cubic lattice. Again, though, the efficiency of this covering decreases exponentially with d; one can do much better using the results of Gritzmann [16].

One can also consider the possibility of using more than one generator of order 2. For instance, one could look at Cayley graphs of degree $2d + 2$ obtained by using d unrestricted generators and two generators of order 2.

However, this is not going to be helpful if one wants to construct large undirected Cayley graphs of a given degree and diameter, at least if the diameter is substantially larger than the degree. For instance, suppose that the degree is fixed as $2d + 2$. If one uses d unrestricted generators and two order-2 generators, then the size of the resulting undirected Abelian Cayley graph of diameter k is at most four times the number of points in the d-dimensional shape S_k. (More precisely, by the methods

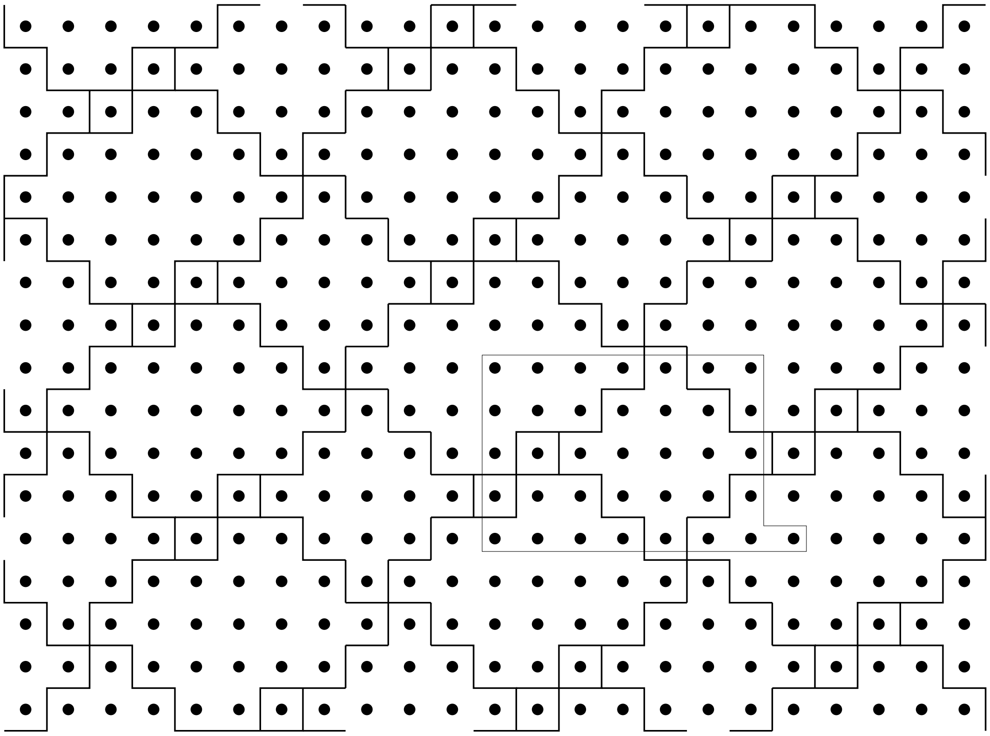


FIG. 11.1. *Lattice covering of $\mathbf{Z}^2$ using S_k and S_{k-1} (shown for $k = 3$).*

TABLE 11.1
Best undirected Cayley graphs of cyclic groups of diameter k using three generators plus one order-2 generator ($3 \le k \le 10$).

k mod 3	0	1	2
a	$2k/3$	$(2k+1)/3$	$(2k-1)/3$
Lattice generators	$(2a, 1, -1)$ $(-1, 2a, -1)$ $(1, 1, 2a)$	$(2a-1, -1, 0)$ $(1, 2a, -1)$ $(0, 1, 2a-1)$	$(2a+1, -1, 0)$ $(1, 2a, -1)$ $(0, 1, 2a+1)$
Extra generator	$(a, a+1, a-1)$	$(a, a, a-1)$	$(a+1, a, a)$
Cyclic group size	$\frac{64k^3+108k}{27}$	$\frac{64k^3+60k-16}{27}$	$\frac{64k^3+60k+16}{27}$
Unrestricted generators	1 $4a^3-2a^2+4a-1$ $4a^3-2a^2+2a-1$	1 $2a-1$ $4a^2-2a+1$	1 $2a+1$ $4a^2+2a+1$
Order-2 generator	$4a^3+3a$	$4a^3-4a^2+3a-1$	$4a^3+4a^2+3a+1$

used above for one order-2 generator, one gets a limit of $|S_k|+2|S_{k-1}|+|S_{k-2}|$.) This limit is $O(k^d)$, which is less than the $O(k^{d+1})$ one gets by using $d+1$ unrestricted generators. The same argument shows that using more than two order-2 generators cannot be useful for large k; one gets larger graphs by replacing two order-2 generators with one unrestricted generator.

If one is interested in the small-diameter case, though (especially when the diameter is less than or equal to the degree), then order-2 generators must be considered. The most extreme version of this would be to make *all* of the generators have order 2. This then becomes precisely the covering radius problem for binary linear codes; see the surveys by Cohen et al. [8, 9] for more on this problem. The resulting graphs would be hypercubes with additional diagonal connections to reduce the diameter.

12. Small-diameter graphs. In most of the previous sections, we considered the case where the degree of the graph (or the number of generators used for a Cayley graph) was a fixed small value, while the diameter bound varied and was usually much larger than the degree. (In particular, the connections with tilings of Euclidean space were useful mainly in this case.) New, interesting problems arise if one instead fixes a (small) diameter bound and allows the degree to vary. (The authors thank the referee for bringing this topic to their attention.)

In this section we will use the variable d to denote the degree of the graph or digraph. In fact, in order to make it easier to state results, we will assume that the set of generators used for undirected Abelian Cayley graphs is closed under negation; hence, d will also be the number of generators used in both directed and undirected cases.

We first dispose of the trivial cases. If the diameter bound is zero, then the graph must have only one vertex. If the diameter bound is one, then the graph must be a complete graph or digraph, and the Moore bound of $d + 1$ is met by such complete graphs. These graphs are Cayley graphs of cyclic groups, using all nonzero elements as generators. (If we were not assuming that the set of generators is closed under negation, then we would only need about half of the nonzero elements as generators in the undirected case. A similar remark applies to all of the other undirected Cayley graphs here.)

Now consider the case of diameter 2 for undirected graphs. The Moore bound for this case is $d^2 + 1$. McKay, Miller, and Širáň [18] have constructed vertex-transitive (but not Cayley) graphs which come relatively close to this bound in a number of cases: For any d of the form $(3q - 1)/2$ where q is a prime power congruent to 1 modulo 4, they construct a vertex-transitive graph of degree d and diameter 2 with $(8/9)(d + \frac{1}{2})^2$ vertices. They note that the best known construction which works for all d, due to Griggs, is the Cayley graph of $\mathbf{Z}_{a+1} \times \mathbf{Z}_{b+1}$, where a and b are $d/2$ rounded up and down, with the generators being those elements having exactly one nonzero coordinate. The number of vertices here is

$$\left\lfloor \frac{d+2}{2} \right\rfloor \left\lceil \frac{d+2}{2} \right\rceil,$$

which can be expressed as $(d^2 + 4d + 4 - \delta)/4$, where δ is 1 for odd d and 0 for even d.

Here is a cyclic variant of the above product construction. For any $a \geq 0$ and $b \geq 0$, the Cayley graph of the cyclic group $\mathbf{Z}_{(2a+1)(b+1)}$ using the $2a + b$ generators

$$\pm 1, \pm 2, \ldots, \pm a, 1(2a + 1), 2(2a + 1), \ldots, b(2a + 1)$$

has diameter 2 (or less in trivial cases). Given a degree d, we get the best results here by splitting $d + 2$ into two parts as nearly equal as possible so that at least one of the

TABLE 12.1
Best undirected Cayley graphs of Abelian groups, diameter 2.

d	n	Generators
1	2	1
2	5	± 1
3	8	$\pm 1, 4$
4	13	± 1, ± 5
5	16	± 1, $\pm 3, 8$
6	21	± 1, ± 2, ± 8
7	26	± 1, ± 2, $\pm 8, 13$

parts is odd, and letting these parts be $2a+1$ and $b+1$. The size of the resulting Cayley graph is $(d^2+4d+4-\delta)/4$, where δ is 1 for odd d, 0 for $d \equiv 0 \pmod 4$, and 4 for $d \equiv 2 \pmod 4$, so it is as good as the product construction 3/4 of the time.

We can get some indication of how close to optimal these constructions are by looking at the results from previous sections of this paper for the diameter-2 case. These give optimal Abelian Cayley graphs for all $d \le 7$, as shown in Table 12.1; all of them turn out to be cyclic. As noted in the preceding section, since we are not in the large-diameter case, we have to consider the possibility that one can get better graphs by using multiple order-2 generators; it turns out that this possibility does not occur in these seven cases.

For $2 \le d \le 7$, the sizes in this table are greater than $(d^2+4d+4)/4$, so the product construction and its cyclic variant are not optimal. Using these specific examples as well as geometric reasoning, we are led to the following two improvements on the cyclic construction.

Construction A. For any $a \ge 0$ and $b \ge 2$, the Cayley graph of the cyclic group $\mathbf{Z}_{(2a+1)(b+2)+1}$, using the $2a+b$ generators

$$\pm 1, \pm 2, \dots, \pm a, a+1+1(2a+1), a+1+2(2a+1), \dots, a+1+b(2a+1),$$

has diameter 2. Given a degree d, we get the best results here by splitting $d+3$ into two parts as nearly equal as possible so that one of the parts is odd while the other is at least 4, and letting these parts be $2a+1$ and $b+2$. For $d \ge 4$, the size of the resulting Cayley graph is $(d^2+6d+13-\delta)/4$, where δ is 1 for even d, 0 for $d \equiv 3 \pmod 4$, and 4 for $d \equiv 1 \pmod 4$.

Construction B. For any $a \ge 2$ and $b \ge 1$, the Cayley graph of the cyclic group $\mathbf{Z}_{(2a+1)(b+1)-4}$, using the $2a+b-2$ generators

$$\pm 1, \pm 3, \pm 4, \dots, \pm a, 1(2a+1)-2, 2(2a+1)-2, \dots, b(2a+1)-2,$$

has diameter 2. (Note that $(j(2a+1)-2)+2 = ((j+1)(2a+1)-2)-(1(2a+1)-2)$.) Given a degree d, we get the best results here by splitting $d+4$ into two parts as nearly equal as possible so that one of the parts is odd and at least 5 while the other is at least 2, and letting these parts be $2a+1$ and $b+1$. For $d \ge 4$, the size of the resulting Cayley graph is $(d^2+8d-\delta)/4$, where δ is 1 for odd d, 0 for $d \equiv 2 \pmod 4$, and 4 for $d \equiv 0 \pmod 4$.

Both Construction A and Construction B give optimal results for $5 \le d \le 7$ (Construction A does for $d = 4$ as well) and do equally well for $d = 8$, but Construction

B does better than Construction A for $d > 8$ (and both of them do better than the product construction). It is possible that Construction B gives the optimal degree-d diameter-2 Abelian Cayley graph for all $d \geq 5$, but it seems more likely that further improvements are possible for large d.

The best one can hope to achieve with an Abelian Cayley graph is the size of the set S_2 (recall that there is a suppressed d here, or actually a $d/2$), which is

$$\binom{d+2}{2} - \left\lceil \frac{d}{2} \right\rceil = d^2/2 + O(d).$$

So one cannot get close to the size of the McKay–Miller–Širáň (MMS) graphs, although the MMS graphs are "almost Abelian Cayley": Šiagiová [21] has shown that each MMS graph is a Abelian lift of a two-vertex graph, while Abelian lifts of one-vertex graphs are just the Abelian Cayley graphs.

The situation for directed Abelian Cayley graphs turns out to be very similar to that in the undirected case. Again one has a product construction (in fact, exactly the same as the product construction for the undirected case) giving a diameter-2 Abelian Cayley digraph on d generators of size $\lfloor (d^2 + 4d + 4)/4 \rfloor$. Also, there is a cyclic version of the construction, using the group $\mathbf{Z}_{(a+1)(b+1)}$ and the $a+b$ generators

$$1, 2, \ldots, a, 1(a+1), 2(a+1), \ldots, b(a+1),$$

which, when $a = \lfloor d/2 \rfloor$ and $b = \lceil d/2 \rceil$, gives a digraph of the same size as that from the product construction (this time there is no exceptional case).

The computations in previous sections do not give as much information for this case; we only have the optimal results for $d = 1, 2, 3$. These are all cyclic, on groups $\mathbf{Z}_3$ (generator 1), $\mathbf{Z}_5$ (generators $1, 2$), and $\mathbf{Z}_9$ (generators $1, 3, 4$), respectively. One can get the following analogues of Constructions A and B, though they are not quite as good.

Construction A'. For any $a \geq 0$ and $b \geq 0$, the Cayley digraph of the cyclic group $\mathbf{Z}_{(a+1)(b+2)-1}$, using the $a + b$ generators

$$1, 2, \ldots, a, a + 1(a + 1), a + 2(a + 1), \ldots, a + b(a + 1),$$

has diameter 2. Given a degree d, we get the best results here by splitting $d + 3$ into two parts as nearly equal as possible, and letting these parts be $a + 1$ and $b + 2$. For $d \geq 4$, the size of the resulting Cayley graph is $\lfloor (d^2 + 6d + 5)/4 \rfloor$.

Construction B'. For any $a \geq 4$ and $b \geq 0$, the Cayley digraph of the cyclic group $\mathbf{Z}_{(a-2)(b+2)+5}$, using the $a + b - 1$ generators

$$1, 3, 4, \ldots, a, a + 1(a - 2), a + 2(a - 2), \ldots, a + b(a - 2),$$

has diameter 2. Given a degree d, we get the best results here by splitting $d + 1$ into two parts as nearly equal as possible so that each of the parts is at least 2, and letting these parts be $a - 2$ and $b + 2$. For $d \geq 3$, the size of the resulting Cayley graph is $\lfloor (d^2 + 2d + 21)/4 \rfloor$.

This time, Construction A' gives the best results for $d \geq 5$, although Construction B' is optimal for $d = 3$ (they are tied when $d = 4$). Again this is a slight improvement over the product construction. Further improvements may be possible, but the best that one can hope for from an Abelian Cayley digraph of diameter 2 is $\binom{d+2}{2}$ vertices, and this again falls far short of the MMS graph (considered as a digraph) or the Moore bound of $d^2 + d + 1$.

One can consider graphs or digraphs of a fixed diameter $k > 2$ and get results similar to those above. In this case the Moore bound is $d^k + O(d^{k-1})$, but the best one can hope for from an Abelian Cayley (di)graph is $d^k/k! + O(d^{k-1})$. An explicit product construction gives $d^k/k^k + O(d^{k-1})$ vertices; improvements on this are possible, but it is not clear whether the constant $1/k^k$ can be improved.

13. Conclusions. We have shown that one can construct Cayley graphs of Abelian groups which have substantially more vertices than traditional toroidal meshes but retain certain desirable features. In particular, routing on the twisted toroidal meshes is easily described in almost the same manner as on toroidal meshes, and the twisted toroidal meshes host the discrete nonperiodic orthogonal grids used in numerical calculations in exactly the same way that toroidal meshes do. In addition, we have shown how our d-dimensional meshes can be constructed with physical wire lengths that remain constant with increasing diameter (and increasing number of vertices) just as the corresponding toroidal meshes can. We have given results which are provably optimal in two dimensions and probably optimal in three dimensions—the physically interesting cases.

In the sequel to this paper, we will show that our methods can be extended to cover certain types of nilpotent groups. These groups yield graphs with cardinalities which still increase polynomially with diameter for a given degree, but with an exponent which is larger than in the Abelian case. One class of groups for which we obtain optimal results includes the groups discussed in Draper and Faber [12]. In particular, we show that the particular groups analyzed in that paper are not optimal for large diameters.

Acknowledgments. We thank Michael Dinneen, whose computer work determining the diameters of many Cayley graphs yielded some of the specific groups listed here and suggested promising families of groups to examine. Thanks also to Francesc Comellas for pointing us to the existing literature on directed loop networks (in particular, the very useful survey paper of Bermond, Comellas, and Hsu [3]), and thanks to Alexander Hulpke for discussions on automorphisms of Abelian groups.

REFERENCES

[1] F. Aguiló, M. A. Fiol, and C. Garcia, *Triple loop networks with small transmission delay*, Discrete Math., 167/168 (1997), pp. 3–16.

[2] F. Annexstein and M. Baumslag, *On the diameter and bisector size of Cayley graphs*, Math. Systems Theory, 26 (1993), pp. 271–291.

[3] J. C. Bermond, F. Comellas, and D. F. Hsu, *Distributed loop computer networks: A survey*, J. Parallel Distrib. Comput., 24 (1994), pp. 2–10.

[4] F. T. Boesch and J.-F. Wang, *Reliable circulant networks with minimum transmission delay*, IEEE Trans. Circuits and Systems, 32 (1985), pp. 1286–1291.

[5] S. Chen and X.-D. Jia, *Undirected loop networks*, Networks, 23 (1993), pp. 257–260.

[6] F. R. K. Chung, *Diameters and eigenvalues*, J. Amer. Math. Soc., 2 (1989), pp. 187–196.

[7] F. R. K. Chung, V. Faber, and T. A. Manteuffel, *An upper bound on the diameter of a graph from eigenvalues associated with its Laplacian*, SIAM J. Discrete Math., 7 (1994), pp. 443–457.

[8] G. D. Cohen, M. G. Karpovsky, H. F. Mattson, Jr., and J. R. Schatz, *Covering radius—survey and recent results*, IEEE Trans. Inform. Theory, 31 (1985), pp. 328–343.

[9] G. D. Cohen, S. N. Litsyn, A. C. Lobstein, and H. F. Mattson, Jr., *Covering radius 1985–1994*, Appl. Algebra Engrg. Comm. Comput., 8 (1997), pp. 173–239.

[10] M. Dinneen and P. Hafner, *New results for the degree/diameter problem*, Networks, 24 (1994), pp. 359–367.

[11] R. Dougherty and H. Janwa, *Covering radius computations for binary cyclic codes*, Math. Comp., 57 (1991), pp. 415–434.

[12] R. Draper and V. Faber, *The Diameter and Mean Diameter of Supertoroidal Networks*, Technical report SRC-TR-90-004, Supercomputing Research Center, Bowie, MD, 1990.

[13] I. Fáry, *Sur la densité des réseaux de domaines convexes*, Bull. Soc. Math. France, 78 (1950), pp. 152–161.

[14] C. M. Fiduccia, R. W. Forcade, and J. S. Zito, *Geometry and diameter bounds of directed Cayley graphs of Abelian groups*, SIAM J. Discrete Math., 11 (1998), pp. 157–167.

[15] R. Forcade and J. Lamoreaux, *Lattice-simplex coverings and the* 84*-shape*, SIAM J. Discrete Math., 13 (2000), pp. 194–201.

[16] P. Gritzmann, *Lattice covering of space with symmetric convex bodies*, Mathematika, 32 (1985), pp. 311–315.

[17] D. Hoylman, *The densest lattice packing of tetrahedra*, Bull. Amer. Math. Soc., 76 (1970), pp. 135–137.

[18] B. D. McKay, M. Miller, and J. Širáň, *A note on large graphs of diameter two and given maximum degree*, J. Combin. Theory Ser. B, 74 (1998), pp. 110–118.

[19] C. Rogers, *Lattice coverings of space*, Mathematika, 6 (1959), pp. 33–39.

[20] G. Sabidussi, *Vertex transitive graphs*, Monatsh. Math., 68 (1964), pp. 426–438.

[21] J. Šiagiová, *A note on the McKay-Miller-Širáň graphs*, J. Combin. Theory Ser. B, 81 (2001), pp. 205–208.

[22] R. Stanton and D. Cowan, *Note on a "square" functional equation*, SIAM Rev., 12 (1970), pp. 277–279.

[23] H. Urakawa, *On the least positive eigenvalue of the Laplacian for the compact quotient of a certain Riemannian symmetric space*, Nagoya Math. J., 78 (1980), pp. 137–152.

[24] C. K. Wong and D. Coppersmith, *A combinatorial problem related to multimodule memory organizations*, J. ACM, 21 (1974), pp. 392–402.

[25] J. L. A. Yebra, M. A. Fiol, P. Morillo, and I. Alegre, *The diameter of undirected graphs associated to plane tessellations*, Ars Combin., 20B (1985), pp. 159–171.

SIAM J. Discrete Math.
Vol. 17, No. 4, pp. 521–540

LOWER BOUNDS ON THE BROADCASTING AND GOSSIPING TIME OF RESTRICTED PROTOCOLS[*]

MICHELE FLAMMINI[†] AND STÉPHANE PÉRENNÈS[‡]

Abstract. In this paper we extend the technique provided in [M. Flammini and S. Pérennès, *Inform. and Comput.*, to appear] to allow the determination of lower bounds on the broadcasting and gossiping time required by the so-called restricted protocols. Informally, a protocol is $(\mathcal{I}, \mathcal{O})$-restricted if at every processor each outgoing activation of an arc depends on at most $\mathcal{I}$ previous incoming activations and any incoming activation influences at most $\mathcal{O}$ successive outgoing activations. Examples of restricted protocols are systolic ones and those running on bounded degree networks.

Thus, under the basic whispering model, we provide the first general lower bound on the gossiping time of d-bounded degree networks in the directed and half-duplex cases. Moreover, significantly improved broadcasting and gossiping lower bounds are obtained for well-known networks such as butterfly, de Bruijn, and Kautz graphs.

All the results are also extended to other communication models such as the c-port and/or postal one.

Key words. broadcasting, gossiping, lower bounding technique, general and specific networks

AMS subject classifications. 68R05, 68R10, 68M10, 68M12

DOI. 10.1137/S0895480101386450

1. Introduction.

Broadcasting (one-to-all communication) and gossiping (all-to-all communication) are well-known communication primitives to disseminate information in communication networks. Such problems have been extensively investigated in recent years for many different networks and under a large variety of models. A survey of the main related results can be found in [9, 8, 5, 12, 10, 11].

We consider first the basic model, called whispering or processor-bound, where at each communication round each processor can have only one active incident link; i.e., the set of the active links forms a matching. Active links are used at the corresponding processors to deliver the items known until that communication round to their neighbors. If the network can be modelled as an undirected graph, it is possible to further distinguish between two different cases: the half-duplex mode, in which active links allow the transmission of messages in only one direction, and the full-duplex mode, in which messages can travel in both directions simultaneously. Clearly such a distinction is meaningless for broadcasting. In fact, for each given root the protocol is fixed in advance and there is a unique item travelling around. Therefore, it traverses each link only once and in one direction.

Starting from the basic whispering model, several generalizations can be defined. First, it is possible to relax the constraint that at each communication round only

[*] Received by the editors March 19, 2001; accepted for publication (in revised form) August 24, 2003; published electronically April 7, 2004. This work was supported by the IST Programme of the EU under contract IST-1999-14186 (ALCOM-FT), by the EU RTN project ARACNE, by the Italian “Progetto Cofinanziato: Allocazione di risorse in reti senza filo,” and by the French MASCOTTE project I3S-CNRS/INRIA/Univ. Nice–Sophia Antipolis.

http://www.siam.org/journals/sidma/17-4/38645.html

[†] Dipartimento di Informatica, Università di L’Aquila, via Vetoio loc.Coppito, I-67010 L’Aquila, Italy (flammini@di.univaq.it).

[‡] MASCOTTE I3S-CNRS/INRIA/Université de Nice–Sophia Antipolis, 2004 Route des Lucioles, BP 93, F-06902 Sophia-Antipolis Cedex, France (Stephane.Perennes@sophia.inria.fr).

one incident link is activated, thus allowing at most $c \geq 1$ active links. This model is called the c-port (see, for instance, [8]). Moreover, in the postal model [1] it is assumed that the communication time can be slower than one time step, so that when items are sent through an active link, they will be available at the arriving processor only after a number of rounds $\delta \geq 1$. The above models are orthogonal in the sense that it is possible to have the c-port also in the postal case. Finally, it is possible to establish some bounds on the lengths of the messages, i.e., on the number of items that can travel simultaneously on a link [3]. However, all the results shown in this paper assume no bound on the messages' length. A more detailed and exhaustive description of the various models together with the corresponding results can be found in [9, 1, 8, 5, 12, 10, 11, 3].

If we restrict our attention to the whispering model, the best lower bounds on the broadcasting time are as follows. Let the parameter d be defined for undirected graphs as the maximum degree minus one and for directed graphs as the maximum out-degree. Then for bounded degree networks in [17, 4] it has been proved that the broadcasting time $b(G)$ of a graph G of n vertices with parameter d satisfies $b(G) \geq \hat{b}(d) \log n$ (from now on all logarithms are assumed to base 2), where $\hat{b}(d) = \frac{1}{\log \varsigma}$ and ς is the largest real number such that $\varsigma^d - \varsigma^{d-1} - \cdots - \varsigma - 1 = 0$. This yields $\hat{b}(2) = 1.4404$, $\hat{b}(3) = 1.1374$, $\hat{b}(4) = 1.0562$, and, for large d, $\hat{b}(d) \approx (1 + \log(e)/2^d)$.

For butterfly and de Bruijn networks (see section 3 for a formal definition) better lower bounds have been obtained in [14] and then improved in [19, 18] by using their structure. For example, in [19] it is proved that for undirected wrapped butterflies, $b(WBF(2, D)) \geq 1.7621D$ ($\approx 1.7621 \log n$) and $b(WBF(3, D)) \geq 2.0002D$ ($\approx 1.2619 \log n$), while for undirected de Bruijn networks, $b(DB(2, D)) \geq 1.4404D$ ($= 1.4404 \log n$) and $b(DB(3, D)) \geq 1.8028D$ ($= 1.1374 \log n$).

Concerning lower bounds on gossiping, in the half-duplex mode there is a general lower bound of $1.4404 \log n$ for all graphs of n vertices [6, 16, 15, 20], this bound being attained for complete graphs. Such a lower bound has been generalized in [7], where it has been shown that any s-systolic (i.e., periodic with period s) gossip protocol in the directed and half-duplex modes for any graph takes at least $\hat{g}(s) \log(n) - O(\log \log n)$ time steps, where $\hat{g}(s) = 1/\log(1/\lambda)$ and λ is the real number between 0 and 1 such that $\sqrt{p_{\lfloor s/2 \rfloor}(\lambda)} \cdot \sqrt{p_{\lceil s/2 \rceil}(\lambda)} = 1$, with $p_j(\lambda) = \lambda + \lambda^3 + \cdots + \lambda^{2j-1}$ for any integer $j > 0$. Moreover, improved lower bounds on the gossiping time of s-systolic protocols are provided in the directed, half-duplex, and full-duplex cases for butterfly, de Bruijn, and Kautz networks. Other results concerning specific networks can be found in [5, 13, 8, 12].

In this paper we extend the technique provided in [7] to allow the determination of lower bounds on the broadcasting and gossiping time required by the so-called restricted protocols. Informally, a protocol has input restriction $\mathcal{I}$ if at every processor the items delivered during any outgoing activation have been communicated to the processor during at most $\mathcal{I}$ of the previous incoming activations. Similarly, the protocol has output restriction $\mathcal{O}$ if at every processor the items received during any incoming activation are delivered by the processor using at most $\mathcal{O}$ successive outgoing activations. A protocol with input restriction $\mathcal{I}$ and output restriction $\mathcal{O}$ is called $(\mathcal{I}, \mathcal{O})$-restricted. As an example, the broadcast protocols running on d-bounded degree networks are $(1, d)$-restricted, the gossip protocols on d-bounded degree networks (∞, d)-restricted, and the s-systolic gossip protocols $(\mathcal{I}, \mathcal{O})$-restricted with $\mathcal{I} + \mathcal{O} \leq s$.

We derive general lower bounds on the broadcasting and gossiping time of restricted protocols. In particular, every $(\mathcal{I}, \mathcal{O})$-restricted protocol for broadcasting or gossiping in the full-duplex mode takes at least $\hat{b}(\mathcal{I},\mathcal{O})\log(n) - O(\log\log n)$ time steps, where $\hat{b}(\mathcal{I},\mathcal{O}) = 1/\log(1/\lambda)$ and λ is the real number between 0 and 1 such that $q_{\mathcal{O}}(\lambda) = 1$, with $q_j(\lambda) = \lambda + \lambda^2 + \cdots + \lambda^j$ for any integer $j > 0$. Similarly, in the directed and undirected half-duplex cases, every $(\mathcal{I},\mathcal{O})$-restricted gossip protocol takes at least $\hat{g}(\mathcal{I},\mathcal{O})\log(n) - O(\log\log n)$ time steps, where $\hat{g}(\mathcal{I},\mathcal{O}) = 1/\log(1/\lambda)$ and λ is the real number between 0 and 1 such that $\sqrt{p_{\mathcal{I}}(\lambda)} \cdot \sqrt{p_{\mathcal{O}}(\lambda)} = 1$, again with $p_j(\lambda) = \lambda + \lambda^3 + \cdots + \lambda^{2j-1}$ for any integer $j > 0$. For $\mathcal{I}$ and $\mathcal{O}$ going to infinity, as a simple corollary this yields the general lower bound independently proved in [6, 16, 15, 20] for all graphs and any (unrestricted) gossip protocol, up to an $O(\log\log n)$ additive factor. Moreover, since as noted above every s-systolic gossip protocol is $(\mathcal{I},\mathcal{O})$-restricted with $\mathcal{I} + \mathcal{O} \leq s$, as a corollary we obtain the same results in [7].

By exploiting the fact that every protocol for a graph with parameter d (defined as above) is (∞, d)-restricted, we then derive the first general lower bound in the directed and half-duplex modes on the gossiping time of networks with fixed parameter d. This gives at least $1.5728\log(n) - O(\log\log n)$ rounds for $d = 2$, $1.4829\log(n) - O(\log\log n)$ rounds for $d = 3$, $1.4555\log(n) - O(\log\log n)$ rounds for $d = 4$, and so forth.

For broadcasting and full-duplex gossiping our results for fixed parameter networks coincide with the lower bounds coming from broadcasting [17, 4] up to an $O(\log\log n)$ additive factor, as it can be easily checked by letting $\lambda = 1/\varsigma$. However, by exploiting more information about the network topology, significantly improved lower bounds are obtained for many relevant interconnection networks such as butterfly, wrapped butterfly, de Bruijn, and Kautz networks (see Figure 4.1). Similarly, new lower bounds for such topologies are also obtained for directed and half-duplex gossip protocols.

Finally, all the results are extended to other communication models, such as the c-port and/or the postal one. Here new bounds are obtained also in the systolic case.

Before concluding the section, let us remark that, in addition to the particular results and numerical values found, a valuable contribution of the paper is the lower bounding technique that extends the ad hoc technique for systolic gossiping presented in [7] in several ways. First, thanks to the use of different matrix norms, it allows us to handle different communication patterns. For instance, all the broadcasting results are completely new. Moreover, a deeper understanding of the intrinsic dependencies of the actions of the protocols and of their effect on the structure of the associated matrices has allowed a general definition of restricted protocols that includes in a unified setting broadcasting ones, systolic ones, and those running on bounded degree graphs. Other examples are protocols performing multicast (many-to-many communication) in which there is a limited number of senders or receivers. Broadcasting is one extremal case. Finally, we have "memoryless" protocols, where every vertex remembers only the items received during a fixed number of previous steps. Such protocols will be further discussed in the conclusive section. Besides these examples, every possible constraint introduced in a protocol in general yields corresponding restrictions that if properly evaluated allow the application of the technique. Concerning the topology-dependent results, even if in this paper we consider only a limited number of networks, the technique has a wide applicability and gives a general framework allowing the determination of nontrivial and improved lower bounds for many other graphs by exploiting very general topological properties. All these considerations hold for different communication models in an orthogonal way. In fact, once a different model is adopted, the whole framework is unchanged and can be exploited once the effect on the protocols' matrices is determined. This outlines the nice basic feature of the technique that all the possible variants (restrictions, communication patterns, topologies, and communication models) are encapsulated in the protocols' matrices and do not affect, or have a very limited effect on, the other details. It is conceivable that other possible variants can be included without substantial effort.

The paper is organized as follows. In the next section we introduce the notation and necessary definitions. In section 3 we give some useful facts and properties on protocols, topologies, matrices, and norms. In section 4 we provide the above mentioned lower bounds on the number of steps needed for broadcasting and full-duplex gossiping. In section 5 we give the lower bounds on the gossiping time in the directed and half-duplex cases for both general and specific networks. In section 6 we extend all the results to the c-port and/or postal models, and finally, in section 7, we give some conclusive remarks and discuss some open questions.

2. Notation and definitions.

Let us first introduce some useful notation and definitions.

We model the network as a digraph $G = (V, A)$ in which vertices represent processors and arcs communication links. Let $n = |V|$ denote the number of vertices in G.

Definition 2.1. *A broadcast (resp., gossip) protocol of length t for $G = (V, A)$ is a sequence $\langle A_1, \ldots, A_t \rangle$ of t subsets $A_1, \ldots, A_t \subseteq A$ subject to the following conditions:*

1. *Each A_i, $1 \leq i \leq t$, is a matching in G (i.e., no two arcs in A_i have a common endpoint).*
2. *For a given root vertex $x \in V$ (resp., for any vertex $x \in V$) and for any other vertex $y \in V$, there exist a path $\langle x_0, x_1, \ldots, x_l \rangle$ with $l \leq t$, $x_0 = x$, and $x_l = y$, and a sequence of positive integers $j_1, \ldots, j_l$ such that $1 \leq j_1 < \cdots < j_l \leq t$ and for every i, $1 \leq i \leq l$, (x_{i-1}, x_i) belongs to A_{j_i}.*

Informally, each A_i represents the set of the arcs that are active at the communication round i. If an arc (x, y) is active at a step i, then at the beginning of step $i + 1$, vertex y knows all the items known by x at the beginning of step i. Then, in order for the sequence of the subsets A_i to be a broadcast (resp., gossip) procedure, starting from the chosen root x (from any vertex x), for any other vertex y there must exist an "informing" path from x to y whose arcs are activated in a proper sequence so that at the end of the protocol y knows the items of x.

Definition 2.2. *Given a digraph G and a vertex $x \in V$, let $b_x(G)$ be the minimum length of a broadcast protocol for G with root x. The broadcasting time of G is $b(G) = \max_{x \in V} b_x(G)$.*

The gossiping time $g(G)$ of a digraph G is the minimum length of a gossip protocol for G.

If we restrict our attention to symmetric digraphs, then the above definition corresponds to half-duplex protocols. In order to obtain the full-duplex case, it is sufficient to slightly modify the condition on the active arcs by saying that at every communication round if (x, y) is active, then (y, x) is also active; i.e., any two active arcs either do not have a common endpoint or are opposite. Clearly such a distinction is meaningless for broadcasting, as there is always a single item travelling around and it traverses each link only once and in one direction.

We denote the activation of an arc (x, y) during round i (i.e., when $(x, y) \in A_i$) as (x, y, i) and denote by $Act(\langle A_1, \ldots, A_t \rangle)$, or simply Act, the set of all the activations of the protocol $\langle A_1, \ldots, A_t \rangle$, i.e., $Act = \{(x, y, i) \mid (x, y) \in A_i, 1 \leq i \leq t\}$. $m = |Act|$ is the total number of activations.

We can then see an informing path from x to y as constituted by the ordered sequence of its activations, i.e., $P(x, y) = \langle(x_0, x_1, j_1), (x_1, x_2, j_2), \ldots, (x_{l-1}, x_l, j_l)\rangle$ with $x_0 = x$, $x_l = y$, and $1 \le j_1 < \cdots < j_l \le t$.

Let $\mathcal{P}(\langle A_1, \ldots, A_t\rangle)$ or simply $\mathcal{P}$ be the set of all the information paths of the protocol. Then the protocol performs a broadcasting from a given root $x \in V$ if $\mathcal{P}$ contains at least one informing path from x to any other vertex $y \in V$. Similarly, for gossiping there must be at least one path for any possible pair of vertices $x, y \in V$.

Given any activation $(x, y, i) \in \mathit{Act}$ incoming at a vertex $y \in V$, in some cases not all the successive activations (y, z, j) outgoing from y depend on (x, y, i), i.e., carry items that have been communicated to y through (x, y, i). It is then possible to define a dependence function $D : \mathit{Act} \to 2^{\mathit{Act}}$ such that, for any $(x, y, i) \in \mathit{Act}$, $D(x, y, i) \subseteq \{(y, z, j) \mid j > i\}$ is the subset of the successive activations outgoing from y that depend on (x, y, i).

Starting from D, let $\mathcal{P}_D \subseteq \mathcal{P}$ be the set of the informing paths in $\mathcal{P}$ that respect D, i.e., such that if $\langle(x_0, x_1, j_1), (x_1, x_2, j_2), \ldots, (x_{l-1}, x_l, j_l)\rangle \in \mathcal{P}_D$, then for any h, $1 \le h < l$, $(x_h, x_{h+1}, j_{h+1}) \in D(x_{h-1}, x_h, j_h)$.

Definition 2.3. *A dependence function D is feasible for a broadcast (resp., gossip) protocol $\langle A_1, \ldots, A_t\rangle$ if given the root $x \in V$ (resp., for any vertex $x \in V$), and given any other vertex $y \in V$, there exists at least one informing path $P(x, y)$ from x to y in $\mathcal{P}_D$.*

Hence, D is feasible if respecting its dependence relationships does not affect the broadcasting (resp., gossiping) property of the protocol, thus not making necessary an increase of its length.

Definition 2.4. *A broadcast (resp., gossip) protocol $\langle A_1, \ldots, A_t\rangle$ is $(\mathcal{I}, \mathcal{O})$-restricted at a given vertex $y \in V$ if it admits a feasible dependence function D that satisfies the following two conditions:*

1. *For any activation $(y, z, j) \in \mathit{Act}$ outgoing from y, $|\{(x, y, i) | (y, z, j) \in D(x, y, i)\}| \le \mathcal{I}$.*
2. *For any activation $(x, y, i) \in \mathit{Act}$ incoming in y, $|\{(y, z, j) | (y, z, j) \in D(x, y, i)\}| \le \mathcal{O}$.*

A broadcast (resp., gossip) protocol is $(\mathcal{I}, \mathcal{O})$-restricted if it is $(\mathcal{I}, \mathcal{O})$-restricted at every vertex $y \in V$.

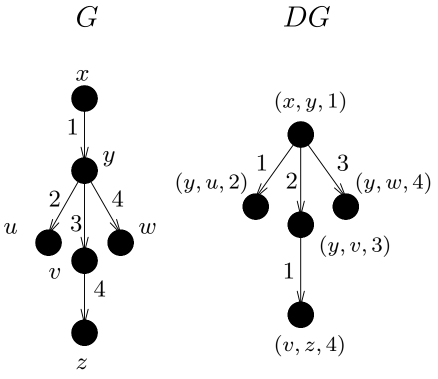
Informally, if the protocol is $(\mathcal{I}, \mathcal{O})$-restricted at a given vertex $y \in V$, then each outgoing activation of y depends on at most $\mathcal{I}$ previous incoming activations and each incoming activation influences at most $\mathcal{O}$ successive outgoing activations.

In order to prove the lower bounds, we now introduce the notion of *delay digraph* of a restricted protocol.

Definition 2.5. *The delay digraph $DG(A_1, \ldots, A_t)$, or simply DG, of an $(\mathcal{I}, \mathcal{O})$-restricted protocol $\langle A_1, \ldots, A_t\rangle$ for G is a weighted digraph $DG = (\mathit{Act}, A')$ with $A' = \{((x, y, i), (y, z, j)) \mid (x, y, i) \in \mathit{Act}, (y, z, j) \in \mathit{Act}, (y, z, j) \in D(x, y, i)\}$ and weight function $\delta((x, y, i), (y, z, j)) = j - i$.*

In DG, given two activations $(x, y, i) \in \mathit{Act}$ and $(y, z, j) \in \mathit{Act}$, $\delta((x, y, i), (y, z, j)) = j - i$ is the delay encountered by an item passing (x, y) at time i to cross (y, z) at time j. Delays are represented only between activations that are dependent according to the dependence function D.

Definition 2.6. *Given an $(\mathcal{I}, \mathcal{O})$-restricted protocol $\langle A_1, \ldots, A_t\rangle$ for G with delay digraph DG and a strictly positive real number $\lambda < 1$, the* delay matrix *$M^{DG}(\lambda)$, or simply $M(\lambda)$, of G with respect to the protocol is the $m \times m$ matrix such that $M(\lambda)_{(x,y,i),(y,z,j)} = \lambda^{\delta((x,y,i),(y,z,j))}$ if $((x, y, i), (y, z, j)) \in A'$, or else $M(\lambda)_{(x,y,i),(y,z,j)} = 0$.*



$M(\lambda)$

$$\begin{array}{c|ccccc|}
 & (x,y,1) & (y,u,2) & (y,v,3) & (y,w,4) & (v,z,4) \\
(x,y,1) & 0 & \lambda & \lambda^2 & \lambda^3 & 0 \\
(y,u,2) & 0 & 0 & 0 & 0 & 0 \\
(y,v,3) & 0 & 0 & 0 & 0 & \lambda \\
(y,w,4) & 0 & 0 & 0 & 0 & 0 \\
(v,z,4) & 0 & 0 & 0 & 0 & 0
\end{array}$$

FIG. 2.1. *An example of broadcast protocol with the associated delay digraph and delay matrix. For the sake of simplicity in G only activated arcs are represented together with their communication rounds. The protocol is $(1,3)$-restricted, since there are at most three entries different from 0 in each row and at most one in each column.*

In the matrix $M(\lambda)$ each row corresponds to an incoming activation and each column to an outgoing one. Since the protocol is $(\mathcal{I},\mathcal{O})$-restricted, there are at most $\mathcal{O}$ entries different from 0 per row and at most $\mathcal{I}$ per column.

A simple example of restricted broadcast protocol is depicted in Figure 2.1, together with the associated delay digraph DG and delay matrix $M(\lambda)$.

Let us finally establish the key property of $M(\lambda)$.

Consider the matrix $M(\lambda)^2 = M(\lambda)\cdot M(\lambda)$. Given any two activations (x,y,i), $(w,z,j) \in Act$, the entry of $M(\lambda)^2$ at row (x,y,i) and column (w,z,j) is given by a sum of products, each corresponding to an intermediate activation $(y,z,k) \in Act$ such that $((x,y,i),(y,w,k)) \in A'$ and $((y,w,k),(w,z,j)) \in A'$. Each such product is equal to $\lambda^{\delta((x,y,i),(y,w,k))}\lambda^{\delta((y,w,k),(w,z,j))} = \lambda^{(k-i)+(j-k)} = \lambda^{j-i}$, so that $(M(\lambda))^2_{(x,y,i),(w,z,j)} = h\lambda^{j-i}$, where h is the number of dipaths from (x,y,i) to (w,z,j) in DG consisting of two arcs.

Generalizing the above argument, for any positive integer l, $(M(\lambda))^l_{(x,y,i),(w,z,j)} = h\lambda^{j-i}$, where h is the number of dipaths from (x,y,i) to (w,z,j) in DG consisting of exactly l arcs. Since all weights in DG are at least equal to 1, any dipath from (x,y,i) to (w,z,j) has at most $j-i<t$ arcs. Therefore, if there exists a dipath from (x,y,i) to (w,z,j) in DG, then, as $0<\lambda<1$,

$$M(\lambda)_{(x,y,i),(w,z,j)} + (M(\lambda))^2_{(x,y,i),(w,z,j)} + \cdots + (M(\lambda))^t_{(x,y,i),(w,z,j)} \geq \lambda^{j-i} > \lambda^t.$$

3. Preliminaries.

In this section we introduce some basic properties and facts that will be used in the following sections.

First, let us give some examples of $(\mathcal{I},\mathcal{O})$-restricted protocols. The first one concerns broadcast protocols, in which it is possible to assume that each vertex has only one incoming activation.

LEMMA 3.1. *Every broadcast protocol is $(1,\mathcal{O})$-restricted.*

Proof. The lemma trivially holds by observing that only the first incoming activation $(x,y,i) \in Act$ at each vertex $y \in V$ influences all the outgoing activations. More precisely, it is possible to put all the outgoing activations (y,z,j) with $j>i$ only in the dependence set $D(x,y,i)$. This clearly maintains in $\mathcal{P}_D$ at least one informing path from the root to any other vertex. □

Protocols running on bounded degree networks are also restricted. In fact, let the parameter d be defined for undirected graphs as the maximum degree minus one and for directed graphs as the maximum out-degree. Then the following lemma holds.

LEMMA 3.2. *Every protocol for a network with fixed parameter d is (∞, d)-restricted.*

Proof. The claim derives by observing that an item of information needs to traverse a given arc only once. Since at any vertex $y \in V$ for every activation (x, y, i) there are at most d successive activations of the type (y, z, j) with $j > i$ corresponding to different arcs, then it is possible to put in the dependence set $D(x, y, i)$ the at most d elements obtained by the first activation after round i of every arc leaving y (clearly except (y, x)). This maintains in $\mathcal{P}_D$ at least one informing path between each pair of vertices having an informing path in $\mathcal{P}$. □

Combining Lemmas 3.1 and 3.2, it is easy to derive that every broadcast protocol for a network with fixed parameter d is $(1, d)$-restricted.

As a last example of restriction we have the s-systolic protocols, i.e., with $A_i = A_{i+s}$ for each $1 \leq i \leq t - s$.

LEMMA 3.3. *Every s-systolic protocol at each vertex is $(\mathcal{I}, \mathcal{O})$-restricted with $\mathcal{I} + \mathcal{O} \leq s$.*

Proof. Given a vertex $y \in V$, let l and r be the number of incoming and outgoing activations, respectively, during a systolic period. Since a period contains s activations, $l + r \leq s$. Analogously as in the previous lemma, since for any activation (x, y, i) there are at most r successive activations of the type (y, z, j) with $j > i$ corresponding to different arcs, it is possible to put in the dependence set $D(x, y, i)$ the r outgoing activations that are within the next systolic period. This maintains in $\mathcal{P}_D$ at least one informing path between each pair of vertices having an informing path in $\mathcal{P}$. The lemma follows by observing that $\mathcal{O} = r$ and that, by construction, each outgoing activation belongs to the dependence set of the l incoming activations occurring in the previous systolic period. Hence $\mathcal{I} = l$, and $\mathcal{I} + \mathcal{O} = l + r \leq s$. □

Notice that in bounded degree networks it is not possible to bound the input restriction. In fact, any outgoing activation might depend on multiple previous activations of the same incoming arc, because each might carry different items. Even if this is also true for systolic protocols, by the periodic fashion the outgoing activation of each arc is repeated every s steps and thus is influenced only by the incoming activations occurring during the previous period. This allows us to also bound the input restriction in such a way as to guarantee that the proper information paths are maintained. In fact, they are all the ones in which any two adjacent activations along the path are distant at most s steps, and this clearly does not increase the dissemination time.

As we will see in the following sections, better lower bounds can be determined when some information about the topology of the network is known. More precisely, this is possible for classes or families of networks containing a large number of vertices distant from the root or among themselves.

DEFINITION 3.4. *Given a family $\mathcal{G}$ of arbitrarily large digraphs and two positive real numbers α and l, $\mathcal{G}$ has an $\langle \alpha, l \rangle$-broadcast separator if, for every digraph $G = (V, A) \in \mathcal{G}$ of n vertices, there exist a root vertex $x \in V$ and a subset of vertices $V_x \subset V$ such that* $\min_{y \in V_x} dist_G(x, y) = l \log(n)(1 - o(1))$ *and* $|V_x| = 2^{\alpha l \log(n)(1-o(1))}$.

An analogous definition can be given for gossiping.

DEFINITION 3.5 (see [7]). *Given a family $\mathcal{G}$ of arbitrarily large digraphs and two positive real numbers α and l, $\mathcal{G}$ has an $\langle \alpha, l \rangle$-gossip separator if, for every* *digraph* $G = (V, A) \in \mathcal{G}$ *of* n *vertices, there exist two subsets of vertices* $V_1 \subset V$ *and* $V_2 \subset V$ *such that* $\min_{x \in V_1, y \in V_2} dist_G(x, y) = l \log(n)(1 - o(1))$ *and* $\min(|V_1|, |V_2|) = 2^{\alpha l \log(n)(1-o(1))}$.

Notice that in the above two definitions α and l depend on the family $\mathcal{G}$ and not on the single digraphs in $\mathcal{G}$. In particular, for every $G \in \mathcal{G}$, α and l are not a function of the number of vertices of G. Moreover, by definition the inequality $\alpha \cdot l \leq 1$ always holds. Clearly, an $\langle \alpha, l \rangle$-gossip separator for G implies also the existence of an $\langle \alpha, l \rangle$-broadcast separator.

In the following, when dealing with digraphs G whose corresponding families $\mathcal{G}$ are clear from the context, for the sake of brevity we will often identify $\mathcal{G}$ simply by G. So, for instance, we will say that G has an $\langle \alpha, l \rangle$-broadcast or gossip separator to mean that $\mathcal{G}$ has such a separator.

The following networks will be considered in what follows.

A *butterfly digraph* of degree d and dimension D, denoted by $BF(d, D)$, has as vertices the $(D + 1)d^D$ tuples $(x, l) \in \{1, \ldots, d\}^D \times \{0, \ldots, D\}$, where $x = x_{D-1}x_{D-2} \ldots x_1x_0$ is a string of length D over $\{1, \ldots, d\}$ and $l \in \{0, \ldots, D\}$ is an integer called level. A vertex $(x_{D-1}x_{D-2} \ldots x_1x_0, l)$ with $l > 0$ is joined with pairwise opposite arcs to the d vertices $(x_{D-1} \ldots x_l, \alpha, x_{l-2}, \ldots x_0, l - 1)$ such that $\alpha \in \{1, \ldots, d\}$.

A *wrapped butterfly digraph* of degree d and dimension D, denoted by $W\vec{B}F(d, D)$, has as vertices the Dd^D tuples $(x, l) \in \{1, \ldots, d\}^D \times \{0, \ldots, D - 1\}$, where $x = x_{D-1}x_{D-2} \ldots x_1x_0$ is a string of length D over $\{1, \ldots, d\}$ and $l \in \{0, \ldots, D - 1\}$ is an integer called level. A vertex $(x_{D-1}x_{D-2} \ldots x_1x_0, l)$ with $l > 0$ has an arc toward the d vertices $(x_{D-1} \ldots x_l\alpha x_{l-2} \ldots x_0, l - 1)$ such that $\alpha \in \{1, \ldots, d\}$ and each vertex $(x_{D-1}x_{D-2} \ldots x_1x_0, 0)$ has an arc toward the d vertices $(\alpha x_{D-2} \ldots x_1x_0, D - 1)$ with $\alpha \in \{1, \ldots, d\}$. The corresponding undirected graph obtained by adding the opposite of each arc is denoted as $WBF(d, D)$ and is generally called a wrapped butterfly graph.

A *de Bruijn digraph* of degree d and dimension D, denoted by $\vec{DB}(d, D)$, has as vertices all the d^D strings of length D over $\{1, \ldots, d\}$. Any vertex $x_{D-1}x_{D-2} \ldots x_1x_0$ has an arc toward the d vertices $x_{D-2}x_{D-3} \ldots x_1x_0\alpha$ such that $\alpha \in \{1, \ldots, d\}$. The corresponding undirected graph, denoted as $DB(d, D)$, is called a de Bruijn graph.

A *Kautz digraph* of degree d and dimension D, denoted by $\vec{K}(d, D)$, has as vertices all the $(d + 1)d^{D-1}$ strings $x_{D-1}x_{D-2} \ldots x_1x_0$ of length D over $\{1, \ldots, d + 1\}$ such that for any j, $0 \leq j \leq D - 2$, $x_j \neq x_{j+1}$. Any vertex $x_{D-1}x_{D-2} \ldots x_1x_0$ has an arc toward the d vertices $x_{D-2}x_{D-3} \ldots x_1x_0\alpha$ with $\alpha \in \{1, \ldots, d + 1\}$ and $\alpha \neq x_0$. The corresponding undirected graph, denoted as $K(d, D)$, is called a Kautz graph.

The families of the butterfly, de Bruijn, and Kautz networks with fixed degree d have large separators.

LEMMA 3.6 (see [7]). *There exists an* $\langle \alpha, l \rangle$*-gossip separator with*

1. $\alpha = \log(d)/2$ *and* $l = 2/\log d$ *for* $BF(d, D)$*;*
2. $\alpha = \log(d)/2$ *and* $l = 2/\log d$ *for* $W\vec{B}F(d, D)$*;*
3. $\alpha = 2\log(d)/3$ *and* $l = 3/(2\log d)$ *for* $WBF(d, D)$*;*
4. $\alpha = \log d$ *and* $l = 1/\log d$ *for* $DB(d, D)$*;*
5. $\alpha = \log d$ *and* $l = 1/\log d$ *for* $K(d, D)$.

We conclude the section by recalling some useful definitions and properties about matrices that are well known in linear algebra (see, for instance, [2, 7]).

Let $\mathbb{R}^m$ be the set of all column vectors $\vec{x} = (x_1, \ldots, x_m)^T$ of m real elements. A real function $| \; | : \mathbb{R}^m \to \mathbb{R}$ is called a *norm* if $|\vec{x}| \geq 0$ for every $\vec{x} \in \mathbb{R}^m$, $|\vec{x}| = 0$ if and only if all the m components of $\vec{x}$ are equal to 0, $|a\vec{x}| = abs(a)|\vec{x}|$ for every $a \in \mathbb{R}$ and

$\vec{x} \in \mathbb{R}^m$ ($abs(a)$ being the absolute value of a), and, finally, $|\vec{x}+\vec{y}| \leq |\vec{x}| + |\vec{y}|$ for all $\vec{x}, \vec{y} \in \mathbb{R}^m$.

For any integer $d > 0$, the d-norm of a vector $\vec{x} \in \mathbb{R}^m$ is defined as $|\vec{x}|_d = (\sum_{j=1}^m abs(x_j)^d)^{1/d}$. The d-norm $||M||_d$ of an $n \times m$ real matrix M is $\sup_{\vec{x}\in\mathbb{R}^m, |\vec{x}|_d \neq 0} \frac{|M\vec{x}|_d}{|\vec{x}|_d}$. $|\vec{x}|_2$ and $||M||_2$ are the Euclidean vector and matrix norms, respectively, while $|\vec{x}|_\infty = \max_{j=1}^m abs(x_i)$ and $||M||_\infty = \max_{i=1}^n (\sum_{j=1}^m abs(M_{i,j}))$ are called vector and matrix maximum norms or norms of the maximum.

For every matrix M with nonnegative real elements, the $||M||_d$ matrix norm satisfies the following properties:

1. $||M||_d \geq 0$;
2. $||M||_d = 0 \Rightarrow M = 0$;
3. for all $a \in \mathbb{R}$, $||aM||_d = abs(a)||M||_d$;
4. $M \geq N$ (i.e., $M_{i,j} \geq N_{i,j}$ for all i,j) $\Rightarrow ||M||_d \geq ||N||_d$;
5. $||M+N||_d \leq ||M||_d + ||N||_d$;
6. $||MN||_d \leq ||M||_d \cdot ||N||_d$;
7. if N is obtained from M by row and column permutations, $||N||_d = ||M||_d$;
8. if M is everywhere null except in k subblocks $M_1, \ldots, M_k$ not sharing any row or column, then $||M||_d = \max_{i=1}^k ||M_i||_d$.

DEFINITION 3.7. *Given an $m \times m$ real matrix M, a nonnull column vector $\vec{x} \in \mathbb{R}^m$ is an* eigenvector *for M with* eigenvalue *e if $M\vec{x} = e\vec{x}$. The spectral radius $\rho(M)$ of M is the maximum absolute value of an eigenvalue of M.*

The spectral radius of a matrix M is related to the Euclidean norm of M. In fact, $||M||_2 = \sqrt{\rho(M^T \cdot M)}$, where M^T is the transpose of M, and if M is symmetric, $||M||_2 = \rho(M)$. Moreover, for any natural matrix norm, that is, defined from a vector norm $|\vec{x}|'$ as $\sup_{\vec{x}\in\mathbb{R}^m, |\vec{x}|' \neq 0} \frac{|M\vec{x}|'}{|\vec{x}|'}$, $||M||' \geq \rho(M)$.

DEFINITION 3.8 (see [7]). *Given an $m \times m$ matrix M, a nonnull column vector $\vec{x} \in \mathbb{R}^m$ is a* semieigenvector *for M with* semieigenvalue *e if $M\vec{x} \leq e\vec{x}$.*

LEMMA 3.9 (see [7]). *Given an $m \times m$ nonnegative matrix M and a strictly positive semieigenvector $\vec{x} \in \mathbb{R}^m$ of M with semieigenvalue e, $\rho(M) \leq e$.*

4. Lower bounds for broadcasting.

In this section we provide new lower bounds on the broadcasting time of the $(\mathcal{I}, \mathcal{O})$-restricted protocols. By Lemma 3.1, we can restrict our attention to the case $\mathcal{I} = 1$. For the sake of brevity, $|\ |$ and $||\ ||$ will denote the vector and matrix norms of the maximum, i.e., $|\ | = |\ |_\infty$ and $||\ || = ||\ ||_\infty$. Moreover, for any integer $j > 0$, the polynomial $q_j(\lambda)$ is defined as $q_j(\lambda) = \lambda + \lambda^2 + \cdots + \lambda^j$.

The following theorem establishes lower bounds holding for any network.

THEOREM 4.1. *Let $\langle A_1, \ldots, A_t \rangle$ be a $(1, \mathcal{O})$-restricted broadcast protocol for a digraph $G = (V, A)$, $\mathcal{O} > 1$. Then $t \geq \hat{b}(1, \mathcal{O}) \log(n) - O(\log \log n)$, where $\hat{b}(1, \mathcal{O}) = \frac{1}{\log(1/\lambda)}$ and λ is the real number such that $0 < \lambda < 1$ and $q_{\mathcal{O}}(\lambda) = 1$.*

Proof. Consider the delay digraph $DG = (Act, A')$. Clearly, $m = |Act| \leq tn/2$, as every node in G can have at most t incident activations—one per round. Let $M(\lambda)$ be the $m \times m$ delay matrix associated to the protocol and x be the root vertex.

Since the protocol performs broadcasting, there exists an informing path from x to every other vertex in G. Let us then choose exactly one such path for every $z \in V$ (clearly, there can be more than one) and let $(x, y, i) \in V'$ and $(w, z, j) \in V'$ be the first and the last activation, respectively, of the path. Recalling the key property of the delay matrix outlined at the end of section 2,

$$M(\lambda)_{(x,y,i),(w,z,j)} + (M(\lambda))^2_{(x,y,i),(w,z,j)} + \cdots + (M(\lambda))^t_{(x,y,i),(w,z,j)} \geq \lambda^{j-i} > \lambda^t. \quad (4.1)$$

Let N be the $m \times m$ boolean matrix N representing the above choices, i.e., such that for every $z \in V$ with $z \neq x$, the element of N in the row of (x, y, i) and column of (w, z, j) is equal to 1 if and only if (x, y, i) and (w, z, j) are the first and the last activation, respectively, of the chosen informing path from x to z. Then, inequality (4.1) can be extended to include all the vertices $z \in V$ in a compact form as

$$M(\lambda) + M(\lambda)^2 + \cdots + M(\lambda)^t > \lambda^t N.$$

By the norm properties,

$$\text{(4.2)} \quad ||M(\lambda)|| + ||M(\lambda)||^2 + \cdots + ||M(\lambda)||^t \geq ||M(\lambda)|| + ||M(\lambda)^2|| + \cdots + ||M(\lambda)^t||$$
$$\geq ||M(\lambda) + M(\lambda)^2 + \cdots + M(\lambda)^t|| > ||\lambda^t N|| = \lambda^t ||N||.$$

Each row of $M(\lambda)$ is associated to an activation (y, z, i) incoming at a vertex z, while the at most $\mathcal{O}$ entries (z, w, j) different from 0 in the row correspond to the successive activations outgoing from z that are influenced by (y, z, i). Each such outgoing activation must belong to a different round and in turn has a different delay, thus corresponding to an entry λ^a with $a > 0$ different from all the other entries in the same row. Hence,

$$\text{(4.3)} \qquad ||M(\lambda)|| = \max_{i=1}^{m} \sum_{j=1}^{m} M_{i,j} \leq \lambda^{a_1} + \lambda^{a_2} + \cdots \lambda^{a_{\mathcal{O}}} \leq q_{\mathcal{O}}(\lambda) = 1$$

by the choice of λ. Moreover,

$$\text{(4.4)} \qquad ||N|| = \max_{i=1}^{m} \sum_{j=1}^{m} N_{i,j} \geq \frac{n-1}{t},$$

since there are $n-1$ entries equal to 1 in N, distributed on at most t rows corresponding to the activations (x, y, i) outgoing from the root.

By (4.2) and (4.3),

$$t \geq ||M(\lambda)|| + ||M(\lambda)||^2 + \cdots + ||M(\lambda)||^t > \lambda^t ||N||,$$

so using (4.4), $t > \frac{\log(||N||) - \log t}{\log(1/\lambda)} \geq \frac{\log(n-1) - 2\log t}{\log(1/\lambda)}$.

Then, if $t \geq \hat{b}(1, \mathcal{O}) \log n$, the claim trivially holds; otherwise it follows from the latter inequality by observing that $2 \log t = O(\log \log n)$. □

As an application of the previous theorem and of Lemma 3.2, general lower bounds can be obtained for bounded degree networks that coincide with the ones provided in [17, 4] up to an $O(\log \log n)$ additive factor. However, if more information about the separating properties of the network is known, by refining the above theorem better bounds can be determined.

THEOREM 4.2. *Let $\langle A_1, \ldots, A_t \rangle$ be a $(1, \mathcal{O})$-restricted broadcast protocol for a digraph $G = (V, A)$ with an $\langle \alpha, l \rangle$-broadcast separator. Then $t \geq \hat{b}(1, \mathcal{O}) \log(n)(1 - o(1))$, where $\hat{b}(1, \mathcal{O}) = \max_{\lambda \,|\, 0<\lambda<1,\, q_{\mathcal{O}}(\lambda) \leq 1} l \frac{\alpha - \log q_{\mathcal{O}}(\lambda)}{\log(1/\lambda)}$.*

Proof. Consider the delay digraph $DG = (Act, A')$ with $m = |Act| \leq tn/2$, and let $M(\lambda)$ be the $m \times m$ delay matrix associated to the protocol. Moreover, let $d = \min_{z \in V_x} dist_G(x, z)$ and $c = |V_x|$, where x and V_x are the root vertex and the set associated to the $\langle \alpha, l \rangle$-broadcast separator of G, respectively.

Similarly as in the proof of Theorem 4.1, there exists an $m \times m$ boolean matrix N satisfying the following conditions:

- For every vertex $z \in V_x$ of G, there exist exactly two activations $(x, y, i) \in Act$ and $(w, z, j) \in Act$ such that the corresponding element of N in the row of (x, y, i) and column of (w, z, j) is equal to 1, while all the elements not corresponding to such pairs are null.
- $M(\lambda)^{d-1} + \cdots + M(\lambda)^t > \lambda^t N$.

The above conditions state that there exists an informing path from x to every vertex $z \in V_x$ of G. Moreover, as x and z are at distance at least d in G, any such path in DG contains at least $d-1$ different arcs. Therefore, recalling that for every integer l, $M(\lambda)^l$ concerns only paths of l arcs in DG, the summation can be restricted to the matrices from $M(\lambda)^{d-1}$ to $M(\lambda)^t$.

Again, by the choice of λ, $||M(\lambda)|| \leq q_{\mathcal{O}}(\lambda) \leq 1$. Moreover, $||N|| = \max_{i=1}^{m} \sum_{j=1}^{m} N_{i,j} \geq \frac{c}{t}$, since there are c entries equal to 1 in N, distributed on at most t rows corresponding to the activations (x, y, i) outgoing from the root.

By the norm properties,

$$\begin{aligned}(t-d+2)||M(\lambda)||^{d-1} &\geq ||M(\lambda)||^{d-1} + ||M(\lambda)||^{d-2} + \cdots + ||M(\lambda)||^t\\ &\geq ||M(\lambda)^{d-1}|| + ||M(\lambda)^{d-2}|| + \cdots + ||M(\lambda)^t||\\ &\geq ||M(\lambda)^{d-1} + \cdots + M(\lambda)^t|| > ||\lambda^t N|| = \lambda^t ||N|| \geq \lambda^t \frac{c}{t},\end{aligned}$$

so that

$$\begin{aligned}t &> \frac{\log(c) - (d-1)\log(||M(\lambda)||) - \log(t-d+2) - \log t}{\log(1/\lambda)}\\ &\geq \frac{\alpha l \log(n) - l\log(n)\log(q_{\mathcal{O}}(\lambda)) - o(l\log n) - \log(t-d+2) - \log t}{\log(1/\lambda)}.\end{aligned}$$

Therefore, $t \geq l\frac{\alpha - \log q_{\mathcal{O}}(\lambda)}{\log(1/\lambda)}\log(n)(1-o(1))$. In fact, if $t \geq \hat{b}(1, \mathcal{O})\log(n)$, the latter inequality trivially holds; otherwise it derives directly from the previous one by observing that $\log(t-d+2) + \log t \leq 2\log t = O(\log(l \log n))$. □

By applying Theorem 4.2 and Lemma 3.6, the following lower bounds for butterfly, de Bruijn, and Kautz networks can be determined.

COROLLARY 4.3. *The lower bounds on the broadcasting time for $BF(d, D)$, $W\vec{B}F(d, D)$, $WBF(d, D)$, $DB(d, D)$, and $K(d, D)$ in Figure* 4.1 *hold.*

Notice that in the above corollary we have made use of gossip separators since broadcast separators cannot yield better values of α and l. In fact, in all the cases $\alpha \cdot l = 1$ and $l \cdot \log n$ coincides with the diameter up to a negligible additive factor.

5. Lower bounds for gossiping. In this section we provide lower bounds on the gossiping time of the $(\mathcal{I}, \mathcal{O})$-restricted gossip protocols. We consider only the directed and half-duplex cases, since in the full-duplex mode the results coincide with those for broadcasting. We give both lower bounds holding for any network and refined ones that exploit the separating properties. As an application, new lower bounds are determined for bounded degree networks, both for general and specific topologies. For the sake of brevity, in this section $|\ |$ and $||\ ||$ will denote the Euclidean vector and

	Ours	Previous
$\hat{b}(BF(2,D))$	2.2104	2
$\hat{b}(BF(3,D))$	1.5203	1.2618
$\hat{b}(BF(4,D))$	1.2938	1.0058 [17, 4]
$\hat{b}(BF(5,D))$	1.1837	1.0014 [17, 4]
$\hat{b}(W\vec{B}F(2,D))$	2.2200	1.7621 [19]
$\hat{b}(W\vec{B}F(3,D))$	1.5244	1.2619 [19]
$\hat{b}(W\vec{B}F(4,D))$	1.2957	1.0058 [17, 4]
$\hat{b}(W\vec{B}F(5,D))$	1.1847	1.0014 [17, 4]
$\hat{b}(WBF(4,D))$	1.1047	1.0058 [17, 4]
$\hat{b}(WBF(5,D))$	1.0433	1.0014 [17, 4]
$\hat{b}(K(2,D))$	1.3042	1.1374 [17, 4]
$\hat{b}(K(3,D))$	1.0433	1.0254 [17, 4]

FIG. 4.1. *Some lower bounds for broadcasting.* $b(G) \geq \hat{b}(G)\log(n)(1-o(1))$. *The unlisted cases coincide with those in* [17, 4]*, except for wrapped butterflies and de Bruijn with* $d = 2, 3$*, for which better bounds are given in* [19]*. The previous lower bounds for* $BF(2,D)$ *and* $BF(3,D)$ *correspond to the trivial ones given by the respective diameters.*

matrix norms, i.e., $|\ | = |\ |_2$ and $||\ || = ||\ ||_2$. Moreover, for any integer $j > 0$, the polynomial $p_j(\lambda)$ is defined as $p_j(\lambda) = \lambda + \lambda^3 + \cdots + \lambda^{2j-1}$.

The following lemmas for gossiping are analogous to the broadcasting lemmas and can be derived as a simple modification of those in [7]. In fact, even if they are not explicitly referred to as restricted protocols in [7], the proof is similar.

LEMMA 5.1. *Let* $\langle A_1, \ldots, A_t\rangle$ *be an* $(\mathcal{I}, \mathcal{O})$*-restricted gossip protocol for a digraph* $G = (V, A)$. *Then* $t \geq \hat{g}(\mathcal{I}, \mathcal{O})\log(n) - O(\log\log n)$*, where* $\hat{g}(\mathcal{I}, \mathcal{O}) = \frac{1}{\log(1/\lambda)}$ *and* λ *is any real number such that* $0 < \lambda < 1$ *and* $||M(\lambda)|| \leq 1$.

LEMMA 5.2. *Let* $\langle A_1, \ldots, A_t\rangle$ *be an* $(\mathcal{I}, \mathcal{O})$*-restricted gossip protocol for a digraph* $G = (V, A)$ *with an* $\langle\alpha, l\rangle$*-gossip separator. Then* $t \geq \hat{g}(\mathcal{I}, \mathcal{O})\log(n)(1 - o(1))$*, where* $\hat{g}(\mathcal{I}, \mathcal{O}) = \max_{\lambda\,|\,0<\lambda<1,\,||M(\lambda)||\leq 1} l\,\frac{\alpha - \log||M(\lambda)||}{\log(1/\lambda)}$.

As a direct consequence of Lemmas 5.1 and 5.2, the problem of deriving lower bounds on the gossiping time is reduced to the determination of the norm of the matrix $M(\lambda)$ associated to $(\mathcal{I}, \mathcal{O})$-restricted gossip protocols. While this task is more or less trivial in broadcasting with the norm of the maximum (this is why we incorporated it directly in the proof), it is more difficult for the Euclidean norm. We now show how such a norm can be determined by means of successive simplification steps performed on $M(\lambda)$.

Observe first that, by the properties of the matrix norm, the value of $||M(\lambda)||$ is not affected by any row or column permutation of $M(\lambda)$. By the definition of DG, for every vertex x of the initial graph G, all the activations (y, x, i) in DG entering x can be connected only to the activations (x, z, j) outgoing from x. It is then possible to permute the rows of $M(\lambda)$ in such a way that for every x all the activations (y, x, j) in DG correspond to adjacent rows and all the activations (x, z, j) to adjacent columns. The resulting matrix is everywhere null, except in disjoint subblocks $M_{x_1}(\lambda), \ldots, M_{x_n}(\lambda)$ such that, for any $i \neq j$, blocks $M_{x_i}(\lambda)$ and $M_{x_j}(\lambda)$ have no common rows or columns. Informally, each subblock $M_x(\lambda)$ corresponds as above to a vertex x of the initial graph G and reports the delays between its incoming and its outgoing activations.

By the properties of the matrix norm, $||M(\lambda)|| = \max_{x \in V} ||M_x(\lambda)||$; hence in the remaining part of this section we concentrate on the determination of each $||M_x(\lambda)||$. In order to simplify the notation, for a given vertex $x \in V$, we will denote $M_x(\lambda)$ simply as $M(\lambda)$ with the understanding that an upper bound on the norm of the local $M(\lambda)$ is also an upper bound on the norm of the global matrix.

As already observed, $M(\lambda)$ expresses the local protocol occurring around x. Every row of $M(\lambda)$ corresponds to an incoming activation of x, that is, to a vertex (y, x, j) in the delay graph DG, and every column corresponds to an outgoing activation of x, that is, to a vertex (x, y, j) of DG. We implicitly assume that at each round an arc incident to x is activated. In fact, any local matrix not satisfying this property can be obtained from one in which the property is satisfied (which corresponds to a complete local protocol at vertex x) by deleting the rows corresponding to the removed incoming activations and the columns corresponding to the removed outgoing activations. This cannot increase $||M(\lambda)||$ and, since in order to apply Lemmas 5.1 and 5.2 we are interested in determining an upper bound on $||M(\lambda)||$, it does not affect the correctness of our proof. Moreover, we assume that the protocol locally at x starts with an incoming activation and ends with an outgoing activation, since this corresponds to deleting initial columns of 0's and final rows of 0's in $M(\lambda)$, again without affecting its norm.

In order to describe the properties of $M(\lambda)$, we first point out that locally at vertex x it is possible to define two sequences of positive integers $\langle (l_j)_{j=\{1\cdots k\}}, (r_j)_{j=\{1\cdots k\}} \rangle$ such that, starting from the first incoming activation at round 1, the protocol locally at x has l_1 incoming activations (from round 1 to round l_1), then r_1 outgoing activations, then l_2 incoming activations, then r_2 outgoing activations, and so on until the last l_k incoming activations and r_k outgoing activations, where k is a suitable positive integer such that $k \leq \lfloor t/2 \rfloor$. Clearly, $\sum_{j=1}^{k} l_j + r_j \leq t$.

Definition 5.3. *Given the couple of sequences $\langle (l_j)_{j=\{1\cdots k\}}, (r_j)_{j=\{1\cdots k\}} \rangle$ associated with the local protocol at vertex x, the incoming (resp., outgoing)* activation block *j is the set of the successive incoming (resp., outgoing) activations corresponding to l_j (resp., r_j).*

Since permuting the rows and columns of $M(\lambda)$ does not affect $||M(\lambda)||$, we can assume the following ordering of the rows and columns of $M(\lambda)$.

- Rows occur in order of incoming activation block and inside each block in reverse order of round. So, for instance, the first row corresponds to the l_1th incoming activation of block 1 and row l_1 to the first incoming activation of block 1.
- Columns occur in order of outgoing activation block and inside each block this time in order of round. Hence column 1 is associated to the first outgoing activation of block 1 and column r_1 to the last outgoing activation of block 1.

By construction, $M(\lambda)$ can be divided in k^2 blocks $B_{1,1}, \ldots, B_{k,k}$ such that $B_{h,j}$ corresponds to the incoming activation block h and the outgoing activation block j and is given by the intersection of the associated rows and columns (see Figure 5.1). If $j < h$, then $B_{h,j}$ has all entries equal to 0, since each incoming activation in block h influences only the successive outgoing activations, i.e., in the outgoing activation blocks from h to k. Moreover, in each block $B_{h,j}$ with $j \geq h$, any entry is different from 0 if the incoming activation associated to its row influences the outgoing activation associated to its column, according to the dependence function D. In fact, only such delays are represented in the delay graph. Two examples of $M(\lambda)$ can be found in Figure 5.1.

$$\begin{array}{|cc|cc|ccc|c|c|}
\lambda & \lambda^2 & \lambda^4 & \lambda^5 & \lambda^8 & \lambda^9 & \lambda^{10} & \lambda^{14} & \cdots \\
\lambda^2 & \lambda^3 & \lambda^5 & \lambda^6 & \lambda^9 & \lambda^{10} & \lambda^{11} & \lambda^{15} & \cdots \\
\hline
0 & 0 & \lambda & \lambda^2 & \lambda^5 & \lambda^6 & \lambda^7 & \lambda^{11} & \cdots \\
\hline
0 & 0 & 0 & 0 & \lambda & \lambda^2 & \lambda^3 & \lambda^7 & \cdots \\
0 & 0 & 0 & 0 & \lambda^2 & \lambda^3 & \lambda^4 & \lambda^8 & \cdots \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda^2 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda^3 & \cdots \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}$$

$$\begin{array}{|cc|cc|ccc|c|c|}
\lambda & 0 & \lambda^4 & 0 & \lambda^8 & \lambda^9 & 0 & 0 & \cdots \\
0 & \lambda^3 & \lambda^5 & 0 & \lambda^9 & 0 & \lambda^{11} & 0 & \cdots \\
\hline
0 & 0 & \lambda & \lambda^2 & 0 & \lambda^6 & \lambda^7 & 0 & \cdots \\
\hline
0 & 0 & 0 & 0 & \lambda & 0 & \lambda^3 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & \lambda^3 & 0 & \lambda^8 & \cdots \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda^3 & \cdots \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}$$

FIG. 5.1. *Two examples of $M(\lambda)$ with $l_1 = 2$, $r_1 = 2$, $l_2 = 1$, $r_2 = 2$, $l_3 = 2$, $r_3 = 3$, $l_4 = 3$, $r_4 = 1, \dots,$ where we have emphasized blocks $B_{h,j}$ with $1 \leq h, j \leq 4$. The first matrix corresponds to an unrestricted protocol, while the second to a $(3, 4)$-restricted one (at most four nonnull entries per row and three per column).*

The following vectors and matrix operations can be used to suitably express $M(\lambda)$ and its semieigenvectors.

- $\vec{\Lambda}(j) = (1, \lambda, \dots, \lambda^{j-1})^T$.
- For ease of notation, given two column vectors $\vec{x}$ and $\vec{y}$ of h and j components, respectively, we denote as $\vec{x}\vec{y} = (\vec{x}^T\vec{y}^T)^T$ the *vertical concatenation* of $\vec{x}$ and $\vec{y}$, i.e., the column vector of $h+j$ components such that the first h components coincide with the ones of $\vec{x}$ and the last remaining j components coincide with the ones of $\vec{y}$.
- Given two $a \times b$ matrices A and B, $A \bigotimes B$ is the componentwise product operation of A and B, i.e., the matrix such that $(A \bigotimes B)_{h,j} = A_{h,j} \cdot B_{h,j}$.

By construction $M(\lambda)$ can be expressed as the componentwise product $D \bigotimes N(\lambda)$ of two matrices D and $N(\lambda)$ constructed as follows.

D is a boolean matrix in which each element is equal to 1 if and only if the corresponding element in $M(\lambda)$ is nonnull, that is, if the outgoing activation associated to its column belongs to the dependence set of the incoming activation associated to its row. Since the protocol is $(\mathcal{I}, \mathcal{O})$-restricted, there are at most $\mathcal{O}$ elements equal to 1 in each row and at most $\mathcal{I}$ in each column (see Figure 5.2).

$N(\lambda)$ can be seen as the local matrix around vertex x in which there is no input-output restriction; that is, every incoming activation influences all the successive outgoing activations (as, for instance, the first matrix of Figure 5.1).

For any integer j, $0 \leq j \leq k$, let sl_j and sr_j be defined as $\sum_{g=1}^{j} l_g$ and $\sum_{g=1}^{j} r_g$, respectively (hence $sl_0 = sr_0 = 0$). Then block $B_{h,j}$ is given by the intersection of

$$
\left|\begin{array}{cc|cc|ccc|c|c}
1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & \cdots \\
0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & \cdots \\
\hline
0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & \cdots \\
\hline
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & \cdots \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}\right|
$$

FIG. 5.2. *The restriction matrix D associated to the $(3,4)$-restricted protocol of Figure 5.1 (at most four entries equal to 1 per row and three per column).*

$$
\left|\begin{array}{cccccc}
 & \vdots & \vdots & \vdots & \vdots & \\
\cdots & D_{sl_{h-1}+1,sr_{j-1}+1}\lambda^{\delta_{h,j}} & D_{sl_{h-1}+1,sr_{j-1}+2}\lambda^{\delta_{h,j}+1} & \cdots & D_{sl_{h-1}+1,sr_j}\lambda^{\delta_{h,j}+r_j-1} & \cdots \\
\cdots & D_{sl_{h-1}+2,sr_{j-1}+1}\lambda^{\delta_{h,j}+1} & D_{sl_{h-1}+2,sr_{j-1}+2}\lambda^{\delta_{h,j}+2} & \cdots & D_{sl_{h-1}+2,sr_j}\lambda^{\delta_{h,j}+r_j} & \cdots \\
 & \vdots & \vdots & & \vdots & \\
\cdots & D_{sl_h,sr_{j-1}+1}\lambda^{\delta_{h,j}+l_h-1} & D_{sl_h,sr_{j-1}+2}\lambda^{\delta_{h,j}+l_h} & \cdots & D_{sl_h,sr_j}\lambda^{\delta_{h,j}+l_h+r_j-2} & \cdots \\
 & \vdots & \vdots & \vdots & \vdots &
\end{array}\right|
$$

FIG. 5.3. *Block $B_{h,j}$ in $M(\lambda)$.*

rows from $sl_{h-1}+1$ to sl_h and columns from $sr_{j-1}+1$ to sr_j. Moreover, if $h \leq j$, given any two integers a and b such that $sl_{h-1}+1 \leq a \leq sl_h$ and $sr_{j-1}+1 \leq b \leq sl_j$, the element of $M(\lambda)$ at row a and column b (hence belonging to $B_{h,j}$) is $M(\lambda)_{a,b} = D_{a,b}\lambda^{\delta_{h,j}}\lambda^{a-sl_{h-1}-1}\lambda^{b-sr_{j-1}-1}$, where $\delta_{h,j}$ is the number of rounds between the last activation of the incoming activation block h and the first activation of the outgoing activation block j; that is, $\delta_{h,j} = 1 + \sum_{g=h}^{j-1}(r_g + l_{g+1}) = sr_{j-1} - sr_{h-1} + sl_j - sl_h + 1$ (see Figure 5.3). In $N(\lambda)$, block $B_{h,j}$ can be suitably expressed as $B_{h,j} = \lambda^{\delta_{h,j}}\vec{\Lambda}(l_h)(\vec{\Lambda}(r_j))^T$.

Let $\vec{\Lambda} = (\lambda^{x_1} \cdot \vec{\Lambda}(r_1))(\lambda^{x_2} \cdot \vec{\Lambda}(r_2)) \cdots (\lambda^{x_k} \cdot \vec{\Lambda}(r_k))$ and $\vec{\Theta} = (\lambda^{x_1} \cdot \vec{\Lambda}(l_1))(\lambda^{x_2} \cdot \vec{\Lambda}(l_2)) \cdots (\lambda^{x_k} \cdot \vec{\Lambda}(l_k))$, where $x_j = \sum_{g=1}^{j-1}(r_g - l_{g+1}) = sr_{j-1} - sl_j + sl_1$, $1 \leq j \leq k$ (thus $x_1 = 0$).

The vectors $\vec{\Lambda}$ and $\vec{\Theta}$ thus defined have the particular property that, if we multiply each column b of $N(\lambda)$ by the bth element of $\vec{\Lambda}$, then any row a contains different odd powers of λ, multiplied times the ath element of $\vec{\Theta}$. This means that the sum of all the elements of row a, which is equal to the ath element of the vector $N(\lambda) \cdot \vec{\Lambda}$, is at most equal to the ath element of $\vec{\Theta}$ multiplied by $p_\infty(\lambda) = \lim_{j\to\infty} p_j(\lambda)$ (see Figure 5.4). Analogously, the ath element of the vector $M(\lambda) \cdot \vec{\Lambda}$ is at most equal to the ath element of $\vec{\Theta}$ multiplied by $p_{\mathcal{O}}(\lambda)$, as in $M(\lambda)$ there are at most $\mathcal{O}$ nonnull entries per row.

Similar considerations hold multiplying the ath column of $M(\lambda)^T$ by the ath element of $\vec{\Theta}$, so that the bth element of $M(\lambda)^T \cdot \vec{\Theta}$ is at most equal to the bth element of $\vec{\Lambda}$ multiplied by $p_{\mathcal{I}}(\lambda)$.

$$\left|\begin{array}{ccccccccc}
\lambda & \lambda^3 & \lambda^5 & \lambda^7 & \lambda^9 & \lambda^{11} & \lambda^{13} & \lambda^{15} & \cdots \\
\lambda^2 & \lambda^4 & \lambda^6 & \lambda^8 & \lambda^{10} & \lambda^{12} & \lambda^{14} & \lambda^{16} & \cdots \\
0 & 0 & \lambda & \lambda^3 & \lambda^5 & \lambda^7 & \lambda^9 & \lambda^{11} & \cdots \\
0 & 0 & 0 & 0 & \lambda & \lambda^3 & \lambda^5 & \lambda^7 & \cdots \\
0 & 0 & 0 & 0 & \lambda^2 & \lambda^4 & \lambda^6 & \lambda^8 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda^2 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda^3 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda^4 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}\right| \qquad \vec{\Theta} = \left|\begin{array}{c}
1 \\ \lambda \\ \lambda \\ \lambda \\ \lambda^2 \\ \lambda \\ \lambda^2 \\ \lambda^3 \\ \vdots
\end{array}\right| \quad \begin{array}{c}
\le p_\infty(\lambda) \\ \le \lambda p_\infty(\lambda) \\ \le \lambda p_\infty(\lambda) \\ \le \lambda p_\infty(\lambda) \\ \le \lambda^2 p_\infty(\lambda) \\ \le \lambda p_\infty(\lambda) \\ \le \lambda^2 p_\infty(\lambda) \\ \le \lambda^3 p_\infty(\lambda) \\
\end{array}$$

$$\left|\begin{array}{ccccccccc} 1 & \lambda & \lambda & \lambda^2 & \lambda & \lambda^2 & \lambda^3 & \lambda & \cdots \end{array}\right| = \vec{\Lambda}^T$$

FIG. 5.4. *$N(\lambda)$ with the vectors $\vec{\Lambda}$ (transposed) and $\vec{\Theta}$, where all entries of each column b have been multiplied by the bth element of $\vec{\Lambda}$. On the right there is an upper bound on the sum of the elements of each row in the new matrix.*

Therefore, it is possible to prove the following lemma.

LEMMA 5.4. $||M(\lambda)|| \le \sqrt{p_{\mathcal{I}}(\lambda)}\sqrt{p_{\mathcal{O}}(\lambda)}$.

Proof. By Lemma 3.9, it is sufficient to show that $\vec{\Lambda}$ is a semieigenvector of $M(\lambda)^T \cdot M(\lambda)$ with semieigenvalue $p_{\mathcal{I}}(\lambda)p_{\mathcal{O}}(\lambda)$, so that $||M(\lambda)|| = \sqrt{\rho(M(\lambda)^T \cdot M(\lambda))} \le \sqrt{p_{\mathcal{I}}(\lambda)}\sqrt{p_{\mathcal{O}}(\lambda)}$.

By definition, if b is any integer such that $sr_{j-1}+1 \le b \le sr_j$ for a suitable $j > 0$, that is, column b in $M(\lambda)$ corresponds to the outgoing activation block j, then the bth element of the semieigenvector $\vec{\Lambda}$ is $\vec{\Lambda}_b = \lambda^{b-sr_{j-1}-1}\lambda^{sr_{j-1}-sl_j+sl_1} = \lambda^{b-sl_j+sl_1-1}$. Similarly, if a is any integer such that $sl_{h-1}+1 \le a \le sl_h$ for a suitable $h > 0$, that is, row a in $M(\lambda)$ corresponds to the incoming activation block h, then the ath element of vector $\vec{\Theta}$ is $\vec{\Theta}_a = \lambda^{a-sl_{h-1}-1}\lambda^{sr_{h-1}-sl_h+sl_1}$. Thus, the ath element of $M \cdot \vec{\Lambda}$ (i.e., the product of row a of M and $\vec{\Lambda}$) is equal to

$$\sum_{j=h}^{k}\sum_{b=sr_{j-1}+1}^{sr_j} D_{a,b}\lambda^{sr_{j-1}-sr_{h-1}+sl_j-sl_h+1}\lambda^{a-sl_{h-1}-1}\lambda^{b-sr_{j-1}-1}\lambda^{b-sl_j+sl_1-1}$$

$$= \sum_{j=h}^{k}\sum_{b=sr_{j-1}+1}^{sr_j} \lambda^{a-sl_{h-1}-1}\lambda^{sr_{h-1}-sl_h+sl_1}D_{a,b}\lambda^{2(b-sr_{h-1})-1}$$

$$= \lambda^{a-sl_{h-1}-1}\lambda^{sr_{h-1}-sl_h+sl_1}\sum_{j=h}^{k}\sum_{b=sr_{j-1}+1}^{sr_j} D_{a,b}\lambda^{2(b-sr_{h-1})-1}$$

$$= \lambda^{a-sl_{h-1}-1}\lambda^{sr_{h-1}-sl_h+sl_1}\sum_{b=sr_{h-1}+1}^{sr_k} D_{a,b}\lambda^{2(b-sr_{h-1})-1}$$

$$\le \lambda^{a-sl_{h-1}-1}\lambda^{sr_{h-1}-sl_h+sl_1}p_{\mathcal{O}}(\lambda) = p_{\mathcal{O}}(\lambda)\vec{\Theta}_a,$$

since at most $\mathcal{O}$ elements $D_{a,b}$ with fixed a are equal to 1, and thus $M \cdot \vec{\Lambda} \le p_{\mathcal{O}}(\lambda)\vec{\Theta}$.

Param. d	2	3	4	5	6	7	8	9
$\hat{g}(d)$	1.5728	1.4829	1.4555	1.4459	1.4425	1.4412	1.4407	1.4405

FIG. 5.5. *The general lower bounds for different values of the parameter d. $g(G) \geq \hat{g}(d)\log(n) - O(\log\log n)$. For limited d no previous lower bounds are known (except the ones inferred from broadcasting in* [17, 4]*), while for $d = \infty$ the value coincides with the* 1.4404 *in* [6, 16, 15, 20].

Similarly, the bth element of $M^T \cdot \vec{\Theta}$ (again with $sr_{j-1}+1 \leq b \leq sr_j$ for a suitable $j > 0$) is equal to

$$\sum_{h=1}^{j} \sum_{a=sl_{h-1}+1}^{sl_h} D_{a,b}\lambda^{sr_{j-1}-sr_{h-1}+sl_j-sl_h+1}\lambda^{a-sl_{h-1}-1}\lambda^{b-sr_{j-1}-1}\lambda^{a-sl_{h-1}-1}\lambda^{sr_{h-1}-sl_h+sl_1}$$

$$= \sum_{h=1}^{j} \sum_{a=sl_{h-1}+1}^{sl_h} \lambda^{b-sl_j+sl_1-1} D_{a,b}\lambda^{2(a+sl_j-sl_h-sl_{h-1})-1}$$

$$= \lambda^{b-sl_j+sl_1-1} \sum_{h=1}^{j} \sum_{a=sl_{h-1}+1}^{sl_h} D_{a,b}\lambda^{2(a+sl_j-sl_h-sl_{h-1})-1}$$

$$\leq \lambda^{b-sl_j+sl_1-1} p_{\mathcal{I}}(\lambda) = p_{\mathcal{I}}(\lambda)\vec{\Lambda}_b,$$

since at most $\mathcal{I}$ elements $D_{a,b}$ with fixed b are equal to 1.

Thus, $M^T \cdot \vec{\Theta} \leq p_{\mathcal{I}}(\lambda)\vec{\Lambda}$, and

$$M^T \cdot M \cdot \vec{\Lambda} \leq p_{\mathcal{O}}(\lambda) M^T \vec{\Theta} \leq p_{\mathcal{I}}(\lambda) p_{\mathcal{O}}(\lambda)\vec{\Lambda};$$

hence the lemma is proved. □

By Lemmas 5.1, 5.2, and 5.4, the following theorems hold.

THEOREM 5.5. *Let $\langle A_1, \ldots, A_t \rangle$ be an $(\mathcal{I}, \mathcal{O})$-restricted gossip protocol for a digraph $G = (V, A)$. Then $t \geq \hat{g}(\mathcal{I}, \mathcal{O})\log(n) - O(\log\log n)$, where $\hat{g}(\mathcal{I}, \mathcal{O}) = \frac{1}{\log(1/\lambda)}$ and λ is the real number such that $0 < \lambda < 1$ and $\sqrt{p_{\mathcal{I}}(\lambda)}\sqrt{p_{\mathcal{O}}(\lambda)} = 1$.*

THEOREM 5.6. *Let $\langle A_1, \ldots, A_t \rangle$ be an $(\mathcal{I}, \mathcal{O})$-restricted gossip protocol for a digraph $G = (V, A)$ with an $\langle \alpha, l \rangle$-gossip separator. Then $t \geq \hat{g}(\mathcal{I}, \mathcal{O})\log(n)(1 - o(1))$, where $\hat{g}(\mathcal{I}, \mathcal{O}) = \max_{\lambda \,|\, 0<\lambda<1,\, \sqrt{p_{\mathcal{I}}(\lambda)}\sqrt{p_{\mathcal{O}}(\lambda)} \leq 1} l^{\frac{\alpha - \log(\sqrt{p_{\mathcal{I}}(\lambda)}\sqrt{p_{\mathcal{O}}(\lambda)})}{\log(1/\lambda)}}$.*

As a consequence of Theorems 5.5 and 5.6 and Lemmas 3.2 and 3.6, new lower bounds can be determined for general and specific network topologies.

COROLLARY 5.7. *Let $G = (V, A)$ be a digraph with fixed parameter $d > 1$. Then, in the directed and half-duplex cases, $g(G) \geq \hat{g}(d)\log(n) - O(\log\log n)$, where $\hat{g}(d) = \frac{1}{\log(1/\lambda)}$ and λ is the real number such that $0 < \lambda < 1$ and $\sqrt{p_\infty(\lambda)}\sqrt{p_d(\lambda)} = 1$.*

Some numerical bounds arising from Corollary 5.7 are listed in Figure 5.5.

COROLLARY 5.8. *The lower bounds on the gossiping time of any protocol for $BF(d, D)$, $W\vec{B}F(d, D)$, $WBF(d, D)$, $DB(d, D)$, and $K(d, D)$ in Figure* 5.6 *hold.*

6. Extensions and generalizations. We now briefly sketch how our results can be extended to other models. In all cases, our lower bound technique can be used as well, with the difference being that the norm of the matrix associated to the

	Ours	Previous
$\hat{g}(BF(2,D))$	2.4200	2.4193 [7]
$\hat{g}(BF(3,D))$	1.7889	1.7788 [7]
$\hat{g}(W\vec{B}F(2,D))$	2.4280	2.4193 [7]
$\hat{g}(W\vec{B}F(3,D))$	1.7825	1.7788 [7]
$\hat{g}(W\vec{B}F(4,D))$	1.5895	1.5876 [7]
$\hat{g}(W\vec{B}F(5,D))$	1.5071	1.5060 [7]
$\hat{g}(WBF(2,D)$	1.9770	1.9750 [7]
$\hat{g}(WBF(3,D))$	1.5544	1.5538 [7]
$\hat{g}(WBF(4,D))$	1.4591	1.4589 [7]
$\hat{g}(\vec{DB}(2,D))$ and $\hat{g}(\vec{K}(2,D))$	1.6375	1.5876 [7]
$\hat{g}(DB(2,D))$ and $\hat{g}(K(2,D))$	1.5965	1.5876 [7]

FIG. 5.6. *Some improved lower bounds for specific networks.* $g(G) \geq \log(n)(1 - o(1))$. *The unlisted entries coincide with the ones in Figure* 5.5 *or in* [7].

protocol is different. Even if we do not show numerical values, all the results for general and specific topologies can be extended by using the new norm.

Let us consider first the c-port model. In this case the condition that only one incident arc can be active at each round is relaxed by admitting at most a given number $c > 0$ of active arcs. For broadcasting (and full-duplex gossiping) it means that, given any incoming activation at a given vertex, there are still at most $\mathcal{O}$ influenced outgoing activations, where $\mathcal{O}$ is the output restriction. However, not all such outgoing activations have a different delay, since up to c of them can belong to the same round. Hence, if $\mathcal{O} = q \cdot c + r$, $||M(\lambda)||_\infty \leq c\lambda + c\lambda^2 + \cdots + c\lambda^q + r\lambda^{q+1}$.

Similar considerations hold for the directed and half-duplex gossiping. Here each row can contain at most c equal entries, and the same holds for each column. By the $(\mathcal{I}, \mathcal{O})$-restriction, a completely analogous argument shows that $||M(\lambda)|| \leq \sqrt{c\lambda + c\lambda^3 + \cdots + c\lambda^{2q_{\mathcal{I}}-1} + r_{\mathcal{I}}\lambda^{2q_{\mathcal{I}}+1}}\sqrt{c\lambda + c\lambda^3 + \cdots + c\lambda^{2q_{\mathcal{O}}-1} + r_{\mathcal{O}}\lambda^{2q_{\mathcal{O}}+1}}$, where $\mathcal{I} = q_{\mathcal{I}} \cdot c + r_{\mathcal{I}}$ and $\mathcal{O} = q_{\mathcal{O}} \cdot c + r_{\mathcal{O}}$.

In the postal model there is an additional integer parameter $\delta > 0$ such that, once a given item has been sent through a link, it will be available at the arriving vertex to be sent on another link only after δ rounds (hence, in the basic model, $\delta = 1$). This means that an incoming activation cannot influence the outgoing activations before the next δ rounds, and thus all the entries of $M(\lambda)$ are either 0 or λ^a with $a \geq \delta$. By completely analogous observations it is then easy to see that both in broadcasting and in directed and half-duplex gossiping (also under c-port) it is sufficient to multiply the basic norm by $\lambda^{\delta-1}$.

Notice that by using the above norms new lower bounds can be obtained for all the restricted protocols and thus for the bounded degree and also systolic ones.

7. Conclusion.

In this paper we have provided a general technique that extends the one presented in [7] and allows the determination of lower bounds on the broadcasting and gossiping time of the restricted protocols. As a consequence new lower bounds have been determined for bounded degree networks in the general case and for specific topologies. Moreover, as a corollary we obtain the same results for systolic protocols in [7].

As noted in the introduction, another example of restricted protocols are the memoryless ones where, given a fixed number of steps $\delta > 0$, every vertex remembers only the items received during the previous δ steps. This means that an incoming activation influences at most δ successive outgoing activations, and an outgoing activation depends on at most δ previous incoming activations; i.e., the protocol is (δ, δ)-restricted.

We believe that memoryless protocols are not just an artificial example of restricted protocols, but that they might be useful in improving the lower bounds for bounded degree networks. In fact, it seems that in this case the input-output restriction is a property that is too local to allow the determination of the best possible lower bound, which we conjecture to be the one corresponding to a (d, d)-restriction. Locally at each vertex it is possible to establish only an (∞, d)-restriction, but it seems that the best dissemination time is achieved when the vertices have an equal alternance of input and output activations, so that in the average the input restriction is also low. A way to limit the input restriction might be to resort to more global properties. We conjecture, in fact, that in a network with fixed parameter d memoryless protocols with $\delta \approx d$ are able to reach the optimal time up to a negligible additive factor. If this is true, then better lower bounds for bounded degree networks can be easily accomplished.

REFERENCES

[1] A. Bar-Noy and S. Kipnis, *Designing broadcasting algorithms in the postal model for message passing systems*, Math. Systems Theory, 27 (1994), pp. 431–452.

[2] A. Berman and R. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics Appl. Math. 9, SIAM, Philadelphia, 1994.

[3] J. Bermond, L. Gargano, A. Rescigno, and U. Vaccaro, *Fast gossiping by short messages*, SIAM J. Comput., 27 (1998), pp. 917–941.

[4] J. Bermond, P. Hell, A. Liestman, and J. Peters, *Broadcasting in bounded degree graphs*, SIAM J. Disc. Math., 5 (1992), pp. 10–24.

[5] J. de Rumeur, *Communication dans les réseaux de processeurs*, Collection Etudes et Recherches en Informatique, Masson, Paris, 1994. (English version to appear.)

[6] S. Even and B. Monien, *On the number of rounds necessary to disseminate information*, in Proceedings of the First ACM Symposium on Parallel Algorithms and Architectures, Association for Computing Machinery, 1989, pp. 318–327.

[7] M. Flammini and S. Pérennès, *Lower bounds on systolic gossip*, Inform. and Comput., to appear; a preliminary version appears in Proceedings of the 11th IEEE International Parallel Processing Symposium, 1997.

[8] P. Fraigniaud and E. Lazard, *Methods and problems of communication in usual networks*, Discrete Appl. Math., 53 (1994), pp. 79–133.

[9] S. Hedetniemi, S. Hedetniemi, and A. Liestman, *A survey of gossiping and broadcasting in communication networks*, Networks, 18 (1986), pp. 319–349.

[10] J. Hromkovič, R. Klasing, E. Stohr, and H. Wagener, *Gossiping in vertex-disjoint paths mode in d-dimensional grids and planar graphs*, Inform. and Comput., 123 (1995), pp. 17–28.

[11] J. Hromkovič, R. Klasing, W. Unger, and H. Wagener, *Optimal algorithms for broadcast and gossip in the edge-disjoint path modes*, Inform. and Comput., 133 (1997), pp. 1–33.

[12] J. Hromkovič, R. Klasing, B. Monien, and R. Peine, *Dissemination of information in interconnection networks (broadcasting and gossiping)*, in Combinatorial Network Theory, D.-Z. Du and D. F. Hsu, eds., Kluwer Academic Publishers, Norwell, MA, 1995, pp. 125–212.

[13] J. Hromkovič, R. Klasing, D. Pardubská, W. Unger, and H. Wagener, *The complexity of systolic dissemination of information in interconnection networks*, RAIRO Theoret. Inform. Appl., 28 (1994), pp. 303–342.

[14] R. Klasing, B. Monien, R. Peine, and E. Stohr, *Broadcasting in butterfly and de Bruijn networks*, Discrete Appl. Math., 53 (1994), pp. 183–197.

[15] D. W. Krumme, G. Cybenko, and K. N. Venkataraman, *Gossiping in minimal time*, SIAM J. Comput., 21 (1992), pp. 111–139.

[16] R. Labahn and I. Warnke, *Quick gossiping by multi-telegraphs*, in Topics in Combinatorics and Graph Theory, Physica-Verlag, Heidelberg, 1990, pp. 451–458.

[17] A. Liestman and J. Peters, *Broadcast networks of bounded degree*, SIAM J. Disc. Math., 1 (1988), pp. 531–540.

[18] S. Pérennès, *Communications dans les réseaux d'interconnexion*, Ph.D. thesis, Université de Nice-Sophia Antipolis, Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis CNRS URA 1376, Nice, France, 1996.

[19] S. Pérennès, *Lower bounds on broadcasting time of de Bruijn networks*, in Second International Euro-Par Conference, Lecture Notes in Comput. Sci. 1123, Springer-Verlag, New York, 1996, pp. 325–332.

[20] V. Sunderam and P. Winkler, *Fast information sharing in a complete network*, Discrete Appl. Math., 42 (1993), pp. 75–86.

SIAM J. Discrete Math.
Vol. 17, No. 4, pp. 541–547

ON LOCAL VERSUS GLOBAL SATISFIABILITY*

LUCA TREVISAN[†]

Abstract. We prove an extremal combinatorial result regarding the fraction of satisfiable clauses in Boolean conjunctive normal form (CNF) formulae enjoying a locally checkable property, thus solving a problem that has been open for several years.

We then generalize the problem to arbitrary constraint satisfaction problems. We prove a tight result even in the generalized case.

Key words. maximum satisfiability, constraint satisfaction, probabilistic method

AMS subject classification. 68R05

DOI. 10.1137/S0895480197326528

1. Introduction.

We deal with the notion of k-satisfiable conjunctive normal form (CNF) formulae introduced and studied by Lieberherr and Specker [4, 5]. A CNF Boolean formula (from now on referred to as *formula*) is k-satisfiable if any subset of k clauses is satisfiable. For any k, let r_k be the largest real (or, better, the supremum of the set of reals) such that in any k-satisfiable set of m clauses, at least $r_k m$ clauses are simultaneously satisfied. Roughly speaking, r_k somewhat shows how local satisfiability implies (fractional) global satisfiability. It has been known that $r_2 = 2/(1+\sqrt{5}) > .618$ [4] (the inverse of the golden ratio), that $r_3 = 2/3$ [5], and that $\lim_{k\to\infty} r_k \leq 3/4$ [3]. Yannakakis [7] has given simplified proofs of the bounds $r_2 \geq 2/(1+\sqrt{5})$ and $r_3 \geq 2/3$ using the probabilistic method.

To the best of our knowledge, determining the exact value of $\lim_{k\to\infty} r_k$ was still an open question.

Our results. We prove that $\lim_{k\to\infty} r_k = 3/4$. Our proof is constructive: For any $r < 3/4$ we show that a k exists such that given a k-satisfiable formula we can find a probability distribution over its variables in such a way that any clause is satisfied with probability at least r. It thus follows that an assignment satisfying at least a fraction r of clauses must exist. It can even be found in linear time using the greedy algorithm in [7].

We then consider a similar question for general constraint satisfaction problems (CSPs). An instance of a CSP is a set of Boolean predicates (or *constraints*) over Boolean variables. The arity of a constraint is the number of variables it depends on. For a fixed integer h, the hCSP is the restriction of CSP where the arity of the constraints is at most h. Note that if an hCSP instance does not contain identically false constraints, then the random assignment, where each variable is true with probability 1/2, will satisfy at least a fraction 2^{-h} of the constraints. We say that a CSP instance is k-satisfiable if any subset of k constraints is satisfiable. For any integers h and k, we define $r_k^{(h)}$ as the supremum of the reals such that for any k-satisfiable instance of hCSP with m constraints, at least $r_k^{(h)} m$ are satisfiable.

*Received by the editors August 27, 1997; accepted for publication (in revised form) August 11, 2003; published electronically April 7, 2004.

http://www.siam.org/journals/sidma/17-4/32652.html

[†]Computer Science Division, University of California, Berkeley, 615 Soda Hall, Berkeley, CA 94720 (luca@cs.berkeley.edu). Part of this work was done while the author was visiting the IBM T.J. Watson Research Laboratories, Yorktown Heights, NY, in October and November 1995.

We prove $\lim_{k\to\infty} r_k^{(h)} = 2^{1-h}$. For the lower bound, it will be easy to use the probabilistic method to obtain $r_{h+1}^{(h)} \geq 2^{1-h}$. In order to prove the upper bound $r_k^{(h)} \leq 2^{1-h}$ for all k we will need a construction of hypergraphs that generalizes the known construction of graphs with small maximum cut and large girth [1].

Preliminary definitions. A *CNF Boolean formula* (or, simply, a *formula*) is a set $\{C_1, \dots, C_m\}$ of *disjunctive clauses* over a set of *variables* $X = \{x_1, \dots, x_n\}$. A disjunctive clause is a disjunction of *literals* in which each literal is either a variable x_i or a *negated* variable $\neg x_i$. An *assignment* for ϕ is a mapping $\tau : X \to \{\mathsf{true}, \mathsf{false}\}$ that associates a *truth value* with any variable. If l is a literal, then we say that τ *satisfies* l if either $l = x$ and $\tau(x) = \mathsf{true}$ or $l = \neg x$ and $\tau(x) = \mathsf{false}$. If $C = l_1 \vee \ldots \vee l_h$ is a clause, we say that τ satisfies C if τ satisfies l_j for some $j \in \{1, \dots, h\}$. A formula ϕ is *k-satisfiable* [4] if any subset of k clauses of ϕ is satisfiable.

An instance of a CSP is a set $\{C_1, \dots, C_m\}$ of *constraints* over a set of *variables* $X = \{x_1, \dots, x_n\}$. A constraint is a Boolean predicate applied to variables from X. An instance of hCSP (where h is an integer) is an instance of a CSP in which the arity of all the predicates is at most h. We define assignments, satisfiability, and k-satisfiability as formulae, with "clauses" replaced by "constraints" in the definitions.

A *random* assignment is a probability distribution over all the assignments. We will restrict ourselves to random assignments, where each variable is assigned true with a certain probability, independently of the assignments to the other variables (bounded independence would also suffice). Thus a random assignment τ_R is entirely specified by the probabilities $\{p_x\}_{x\in X}$, where $\mathbf{Pr}[\tau_R(x) = \mathsf{true}] = p_x$. To shorten notation, we will write $\mathbf{Pr}[x = \mathsf{true}]$ in place of $\mathbf{Pr}[\tau_R(x) = \mathsf{true}]$ when the random assignment is clear from the context.

2. The CNF result.

2.1. Yannakakis's argument and how to extend it: An informal account. In order to present the main ideas underlying our proof, let us first recall Yannakakis's proof that $r_3 \geq 2/3$. Given a 3-satisfiable formula, he shows how to find a probability distribution over the variables that satisfies all clauses with probability at least $2/3$. If a literal l occurs in a unary clause, then we set $\mathbf{Pr}[l = \mathsf{true}] = 2/3$. Note that this definition is consistent since it is impossible to have the clauses (x) and $(\neg x)$ in the same 3-satisfiable formula. To all the other variables (the ones that do not occur in unary clauses), if any, we give value true with probability $1/2$. Ternary clauses, or longer ones, are satisfied with probability at least $1 - (2/3)^3 = .7037\cdots > 2/3$. It remains to consider binary clauses. If at least one of the literals in a binary clause is true with probability at least $1/2$, then the probability that the clause will be satisfied is at least $1 - (2/3)1/2 = 2/3$. The only bad case happens when both literals are true only with probability $1/3$, but this is impossible because it would mean that the formula contains clauses $(l_1), (l_2), (\neg l_1 \vee \neg l_2)$, which contradicts the fact that it is 3-satisfiable.

When we want to achieve the same construction with an arbitrary $r < 3/4$ in place of $2/3$, we run into some trouble. Let us try with $r = .74$. Literals occurring in unary clauses must be true with probability $.74$. If l occurs in a unary clause, and we have the clause $\neg l \vee x$, then x must be true with probability at least $1-(1-r)/r = .6486\ldots$. Then we have to consider literals occurring with $\neg x$ in a binary clause: they have to be true with probability at least $.5991\ldots$. There are three more cases to be considered (probabilities will be, respectively, $0.566\ldots$, $0.5406\ldots$, and $0.5191\ldots$); we still have to make sure that we are not introducing any inconsistency, and we have to deal with ternary and 4-ary clauses (clauses with 5 or more literals are satisfied with probability at least $1-(.74)^5 > .74$).

The above discussion leaves us with the idea that the range of values for the probabilities of the literals should be $p_1 = r$, $p_2 = 1-(1-r)/r$, $p_3 = 1-(1-r)/p_2$, $\dots$, $p_k = 1-(1-r)/p_{k-1}$. It is comforting that this sequence will eventually go below 1/2, where it can be stopped (Lemma 2).

We also note that, when we want to achieve a ratio close to 3/4, the number of cases to be considered explodes, and that a uniform method to deal with them has to be found.

In order to attribute probabilities to the literals in a uniform way, we introduce the idea of *ranking* them according to the depth of *proofs* of the literals in a simple propositional proof system, whose axioms are the clauses of the formula. This gives at the same time a uniform way to deal with clauses of different lengths and a simple method to show that the assignment of probabilities is consistent.

2.2. The actual proof. The following definition gives the values that we will use in the probability distribution.

DEFINITION 1. *For any real $r \neq 0$, we define the sequence $\{a_i^r\}_{i\geq 1}$ as follows:*

- $a_1^r = r$;
- $a_{i+1}^r = 1-(1-r)/a_i^r$.

If we start from a number $r < 3/4$, the sequence eventually goes below 1/2.

LEMMA 2. *For any r such that $1/2 < r < 3/4$, an $h(r)$ exists such that $a_{h(r)}^r < .5$.*

Proof. Suppose the lemma is false. Note that if $a_i^r > 0$, then $a_{i+1}^r < a_i^r$, as can be easily proved by induction. Then we have a monotonically decreasing sequence that is lower bounded by 0.5: such a sequence must have a limit, so let it be x. Then x is a real root of the equation

$$x = 1-(1-r)/x,$$

that is,

$$x^2 - x + 1 - r = 0.$$

But such an equation has no real root when $1-4(1-r) < 0$, that is, when $r < 3/4$. □

The following definition allows us to *rank* literals and will be used to assign to each of them the right probability.

DEFINITION 3 (provability). *Given a CNF formula ϕ,*

- *if $(l) \in \phi$, then l is 1-provable in ϕ.*
- *if $(l_1 \vee \cdots \vee l_h) \in \phi$ and $\neg l_j$ is i_j-provable in ϕ for $j = 1, \dots, h-1$, then l_h is $(1+\max\{i_1, \dots, i_{h-1}\})$-provable in ϕ.*

A literal is exactly i-provable *in ϕ if i is the smallest integer such that it is i-provable in ϕ.*

Note that in a formula there can be literals that are not i-provable for any i. More generally, a formula may contain a literal l such that, for every i, neither l nor the negation of l is i-provable. For example, if a formula contains no unit clause, then no literal is i-provable for any i. This is no coincidence, since in formulae without a unit it is simple to satisfy at least 3/4 of the clauses. The notion of i-provability helps us identify those literals that create the most “trouble” when we try to satisfy a large fraction of the clauses of a formula.

LEMMA 4. *Let ϕ be a formula with clauses of length at most 4. If x is i-provable in ϕ and $\neg x$ is j-provable in ϕ, then ϕ is not $(3^{i+1} + 3^{j+1} - 2)$-satisfiable.*

Proof. Simple induction shows that when a literal l is i-provable in ϕ, then a set S_l of at most $3^{i+1}-1$ clauses of ϕ exists such that any assignment that satisfies all the clauses in S_l must also satisfy l. The base case is for $i=1$, in which case the set is just $\{(l)\}$, of size $1 < 3^{1+1}-1$. If l is i-provable, then there is a clause $(l \bigvee_{j=1}^{k} l_j)$ with $0 \leq k \leq 3$ such that $\neg l_j$ is i_j-provable with $i_j \leq i-1$. By inductive hypothesis there are sets $S_1, \ldots, S_k$, each of size at most $3^i - 1$ such that each assignment that satisfies the clauses of S_j must also contradict l_j. Let us now take $S_l = \{(l \bigvee_{j=1}^{k} l_j)\} \bigcup_{j=1}^{k} S_j$. Then $|S_l| \leq 3^{i+1}-1$, and an assignment that satisfies all the clauses of S_l must contradict $l_1, \ldots, l_k$ but also satisfy $(l \bigvee_j l_j)$, and therefore it must satisfy l.

The lemma now follows by observing that the set $S_x \cup S_{\neg x}$ has at most $3^{i+1} + 3^{j+1} - 2$ clauses, and no assignment can satisfy all of them. ∎

The next theorem is clearly a sufficient condition for $\lim_{k\to\infty} r_k \geq 3/4$.

THEOREM 5. *For any r such that $1/2 < r < 3/4$, a k exists (depending on r) such that for any k-satisfiable formula ϕ we can find in polynomial time a probability distribution over the variables in such a way that any clause is satisfied with probability at least r.*

Proof. For any variable x, the probability p_x of x to be true will be a rational between r and $1-r$, and, in particular, between $1/4$ and $3/4$. This implies that any 5-ary clause is satisfied with probability at least $1-(3/4)^5 > 3/4$. Thus we only have to care about unary, binary, ternary, and 4-ary clauses. Let us fix $r < 3/4$ and let $k = 2 \cdot 3^{h(r)+1} - 1$. Let ϕ be a k-satisfiable formula, and let ϕ_4 be the subset of clauses of ϕ of length at most 4. Observe that if some literal is i-provable in ϕ_4 for some $i \leq h(r)$, then it is not possible that its complement is j-provable in ϕ_4 for some $j \leq h(r)$.

We shall use the values $a_1^r, \ldots, a_{h(r)-1}^r, 0.5$ in our probability distribution. Let $p_i = a_i^r$ for $i = 1, \ldots, h(r)-1$ and $p_{h(r)} = 1/2$. The probability distribution is as follows.

$$\mathbf{Pr}[x = \mathsf{true}] = \begin{cases} p_i & \text{if } x \text{ is exactly } i\text{-provable in } \phi_4 \text{ for } i \leq h(r)-1, \\ 1-p_i & \text{if } \neg x \text{ is exactly } i\text{-provable in } \phi_4 \text{ for } i \leq h(r)-1, \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$

It should be clear that the definition above is consistent. Recall that the sequence $p_1, \ldots, p_{h(r)}$ is decreasing. So if a variable x is exactly i-provable for some $i < h(r)$, the smaller i is, the larger $\mathbf{Pr}[x = \mathsf{true}]$ is.

CLAIM 6. *Under the probability distribution above, any clause of ϕ is false with probability at most $1-r$.*

Proof. The statement is easy to prove for unary clauses and for clauses with five or more literals.

Let $C = (l_1 \vee \cdots \vee l_h)$ be a clause with two or more literals; we assume $\mathbf{Pr}[l_1 = \mathsf{false}] \leq \mathbf{Pr}[l_1 = \mathsf{false}] \leq \cdots \leq \mathbf{Pr}[l_h = \mathsf{false}]$. If $\mathbf{Pr}[l_2 = \mathsf{false}] \leq 1/2$, then also $\mathbf{Pr}[l_1 = \mathsf{false}] \leq 1/2$ and $\mathbf{Pr}[C \text{ is false}] \leq 1/4 < 1-r$. It remains to consider the case $\mathbf{Pr}[l_2 = \mathsf{false}] > 1/2$. Then $\neg l_2$ is exactly i_2-provable for some $i_2 \leq h(r)-1$; also $\neg l_3$ and $\neg l_4$ (if present) are exactly i_3-provable (resp., i_4-provable) for some $i_3 \leq i_2$ (resp., $i_4 \leq i_2$). It follows that l_1 is exactly i_1-provable for some $i_1 \leq i_2 + 1$, and thus $\mathbf{Pr}[l_1 = \mathsf{false}] = 1 - p_{i_1} \leq 1 - a_{i_1} = (1-r)/a_{i_1-1}$,[1] while $\mathbf{Pr}[l_2 = \mathsf{false}] = p_{i_2} = a_{i_2} \leq$

[1] Note that if $i_1 = h(r)$, then l_1 will be assigned probability 1/2 (that is, exactly p_{i_1}) not because it is exactly $h(r)$-provable, but because it is not i-provable for $i < h(r)$ and, of course, neither is its complement (so l_1 falls in the "otherwise" part of the definition).

a_{i_1-1}. As a consequence, we have

$$\mathbf{Pr}[C \text{ is false }] \leq \mathbf{Pr}[l_1 = l_2 = \mathsf{false}] \leq 1 - r.$$

The theorem thus follows. □

3. Constraint satisfaction problems.

LEMMA 7. *Let ϕ be an $(h+1)$-satisfiable instance of hCSP. Then it is possible to satisfy at least 2^{1-h} of the constraints.*

Proof. We describe a random assignment that satisfies each constraint with probability at least 2^{1-h}.

We say that a constraint is *conjunctive* if there is only one assignment of its variables that satisfies it. For any variable that occurs in a conjunctive constraint, we set it to the value imposed by the constraint. This is consistent (otherwise the instance would not be 2-satisfiable). This partial assignment does not contradict any (nonconjunctive) constraint (otherwise the instance would not be $(h+1)$-satisfiable). We give probability $1/2$ to all the other variables. It is easy to see that any constraint that is not satisfied by the partial assignment is true with probability at least $2/2^h$: indeed, either it is still h-ary and has two or more satisfying assignments, or its arity has been decreased by the partial assignment, and so it is true with probability at least $1/2^{h-1}$. □

Let h, $r < 2^{1-h}$, and k be fixed. We will show how to find a k-satisfiable instance of hCSP such that only a fraction r of its constraints is simultaneously satisfiable.

We will use only one type of constraint, the HYPERCUT^h constraint, defined as follows:

$$\text{HYPERCUT}^h(x_1, \ldots, x_{h-1}, y) \equiv (x_1 \neq y) \wedge (x_1 = \cdots = x_{h-1}).$$

For $h = 2$ this is the xor constraint that gives rise to a CSP that is equivalent to 2-colorability.

For a set ϕ of HYPERCUT^h constraints, if HYPERCUT$^h(x_1, \ldots, x_{h-1}, y) \in \phi$, then we say that, for any $i = 1, \ldots, h-1$, x_i is *adjacent* to y (and that y is adjacent to x_i) in ϕ. A *cycle of length l* $(l \geq 3)$ is a sequence of variables $x_1, \ldots, x_l$ such that x_l is adjacent to x_1 and x_i is adjacent to x_{i+1} for $i = 1, \ldots, l-1$. The reader should be easily convinced that ϕ is satisfiable if and only if it does not contain a cycle of odd length. The next theorem is well known for the case $h = 2$ [1].

LEMMA 8. *For any integers k, h, and any $\epsilon > 0$, there exists a family of m HYPERCUT^h constraints such that no more than $(2^{1-h} + \epsilon)m$ are simultaneously satisfiable and any k of them are satisfiable.*

Sketch of proof. To meet the second requirement we just have to construct an instance without short cycles of odd length. The following construction will work for all sufficiently large n. We fix a (small) constant $\delta > 0$ and a (large) constant c such that

$$\begin{aligned} 2^{1-h}(1+\delta)/(1-2\delta) &< 2^{1-h} + \epsilon, \\ 2k(2c(k-1))^k &\leq \delta c n, \\ c &\geq 6 \log e \log \frac{1}{\delta^2} 2^{h-1}. \end{aligned}$$

Let $m = cn$, and let $s(n) = n\binom{n-1}{h-1}$ be all the possible HYPERCUT^h constraints over the variable set $\{x_1, \ldots, x_n\}$. We construct a random instance of hCSP by choosing each of the $s(n)$ constraints independently with probability $m/s(n)$. We make the following claims:

1. With probability at least .9, the number of constraints in the random instance is at least $m(1-\delta)$.
2. With probability at least .9, the generated instance is such that any assignment satisfies at most $2^{1-h}(1+\delta)m$ constraints.
3. With probability at least .5, there are at most $2k(2c(k-1))^k$ cycles of length $\leq k$ in the generated instance.

With positive probability, a random instance will satisfy all three properties. In particular, there will exist an instance satisfying such properties. By removing from it a constraint for each cycle of length $\leq k$, we obtain a new instance with no cycle of length $\leq k$, $m' \geq m(1-2\delta)$ constraints, and such that no assignment satisfies more than $(2^{1-h}+\epsilon)m'$ constraints. This modified instance proves the lemma.

We now prove the three claims.

1. The average number of constraints is m. By Chernoff bounds, it will be at least $(1-\delta)m$ with probability at least $1-e^{-\delta^2 m/2}$, which is larger than .9 for sufficiently large n.
2. If we fix one of the 2^n possible assignments, which gives value `true` to tn variables, and value `false` to $(1-t)n$, it will satisfy a randomly chosen constraint with probability
$$t^{h-1}(1-t)+(1-t)^{h-1}t \leq (1/2)^{h-1}.$$
From Chernoff bounds, the probability that, for a random instance, there exists an assignment satisfying more than $m2^{1-h}(1+\delta)$ constraints, is at most
$$2^n e^{-\delta^2 2^{1-h} cn/3} \leq 2^{-n} \leq .1$$
for sufficiently large n.
3. There are $n(n-1)\cdots(n-l+1)$ possible cycles of length l. Thus, there are at most kn^k cycles of length $\leq k$. Two fixed variables are adjacent with probability at most $2c(k-1)/n$. For different pairs of variables, their probability of being adjacent is not necessarily independent, but they are negatively correlated, so a given cycle exists with probability at most $(2c(k-1)/n)^k$. The average is at most $k(2c(k-1))^k$; with probability at most .5 the actual number is more than twice the average. □

THEOREM 9. *For any $h \geq 2$, $\lim_{k\to\infty} r_k^{(h)} = 2^{1-h}$.*

4. Conclusions.

It is a startling coincidence that 3/4 is the integrality gap of the tightest known linear programming relaxation of MAX SAT [2] and that 2^{1-h} is the integrality gap of the tighter known linear programming relaxation of MAX hCSP [6]. It would be interesting to understand if this fact has some explanation.

REFERENCES

[1] P. Erdös, *On bipartite subgraphs of graphs*, Mat. Lapok, 18 (1967), pp. 283–288.

[2] M. X. Goemans and D. P. Williamson, *New 3/4-approximation algorithms for the maximum satisfiability problem*, SIAM J. Discrete Math., 7 (1994), pp. 656–666.

[3] M. Huang and K. Lieberherr, *Implications of forbidden structures for extremal algorithmic problems*, Theoret. Comput. Sci., 40 (1985), pp. 195–210.

[4] K. Lieberherr and E. Specker, *Complexity of partial satisfaction*, J. ACM, 28 (1981), pp. 411–422.

[5] K. Lieberherr and E. Specker, *Complexity of Partial Satisfaction* II, Tech. report 293, Department of Electrical Engineering and Computer Science, Princeton University, Princeton, NJ, 1982.

[6] L. Trevisan, *Positive linear programming, parallel approximation, and PCP's*, in Proceedings of the 4th European Symposium on Algorithms, Lecture Notes in Comput. Sci. 1136, Springer-Verlag, Berlin, New York, 1996, pp. 62–75.

[7] M. Yannakakis, *On the approximation of maximum satisfiability*, J. Algorithms, 17 (1994), pp. 475–502.

SIAM J. Discrete Math.
Vol. 17, No. 4, pp. 548–570

ON THE MAXIMAL CODES OF LENGTH 3 WITH THE 2-IDENTIFIABLE PARENT PROPERTY[*]

VU DONG TÔ[†] AND REIHANEH SAFAVI-NAINI[†]

Abstract. A q-ary code has identifiable parent property (IPP) if it allows one of the parents of a descendant word to be found. A 2-IPP code ensures that at least one parent of a pirate word constructed by a coalition of two users can be found. In this paper, we answer a question raised in [H. D. L. Hollmann et al., *J. Combin. Theory Ser. A*, 82 (1998), pp. 121–133] and show that $F(q)$, the maximum number of codewords in a 2-IPP code of length 3, satisfies $|\mathcal{G}_0| \leq F(q) \leq |\mathcal{G}_0| + 2$, where $\mathcal{G}_0$ is a well-defined graph. We also give an efficient algorithm ($O(q^3)$) for finding maximal codes.

Key words. IPP code, frameproof, c-secure code, collusion secure fingerprinting, graph theory, color graph

AMS subject classifications. 68R10, 90C27, 90C47

DOI. 10.1137/S0895480102400424

1. Introduction.

In this paper, we are interested in codes with identifiable parent property (IPP) that allow tracing of illegal copies of digital objects protected by an embedded fingerprint. 2-IPP codes were introduced in [3] and further investigated in [6] and [5]. In [3], bounds on $F(n,q)$, the maximum number of codewords in q-ary 2-IPP codes of length n, were investigated and, for the first nontrivial case, that is, $F(3,q)$, it was shown that $F(3,q) \leq 3q-1$. However, this result is nonconstructive, and codes that achieve the bound with equality are not known. In this paper, we extend this result by giving the construction of a code which, for some values of q, has the maximum number of codewords and, for other values of q, has at most two less codewords compared to the maximal code.

Related works. Tracing pirates was first considered in the context of broadcast encryption [2]. Frameproof codes and c-secure codes were introduced in [1] for protection against illegal copying of software. Traceability codes were studied in [2] and [6]. The relationship between these codes was investigated in [5]. In the rest of this paper, we will concentrate on 2-IPP codes.

Let $\mathcal{A}$ be an alphabet of size q, $|\mathcal{A}| = q$, and $\mathcal{A}^n$ denote the set of n-tuples over $\mathcal{A}$. A code $\mathcal{C}$, of length n and size N over $\mathcal{A}$, is a subset of size N of $\mathcal{A}^n$ and is called an *(N,n,q)-code.*

A *codeword* $c \in \mathcal{C}$ is an n-tuple $(c_1, c_2, \ldots, c_n)$. For a subset $X \subset \mathcal{C}$, we define the set of *descendants* of X as

$$desc(X) = \{a \in \mathcal{A}^n : a_i \in \{x_i : x \in X\},\ 1 \leq i \leq n\}.$$

If $a \in desc(X)$, then $x \in X$ is a *parent* of a. The set of descendants is a subset of $\mathcal{A}^n$ that can be constructed by a coalition of users who have the codewords in X.

[*] Received by the editors January 3, 2002; accepted for publication (in revised form) September 30, 2003; published electronically April 7, 2004. This research was supported in part by the Australian Research Council and Motorola Australian Research Centre.

http://www.siam.org/journals/sidma/17-4/40042.html

[†] School of IT & CS, University of Wollongong, NSW, 2522, Australia (dong@uow.edu.au, rei@uow.edu.au).

For a code $\mathcal{C}$, define

$$desc_w(\mathcal{C}) = \{a \in \mathcal{A}^n : a \in desc(X), X \subset \mathcal{C}, |X| \leq w\}.$$

A *w-IPP code* is a code with the property that for all words in $desc_w(C)$ at least one parent can be found. That is, for any $a \in desc_w(\mathcal{C})$,

$$\bigcap_{\{a \in desc(X), |X| \leq w\}} X \neq \emptyset.$$

2. *q*-ary code graph representation. Let $\mathcal{C}$ be a q-ary code of length 3. We follow [3] and define a 3-color graph $\mathcal{C}^*$ for $\mathcal{C}$ as follows (see Figure 1):

(i) Each node in $\mathcal{C}^*$ corresponds to a codeword in $\mathcal{C}$;

(ii) two nodes in $\mathcal{C}^*$ are joined by an edge of color i if their corresponding codewords in $\mathcal{C}$ have their symbols in the position i equal.

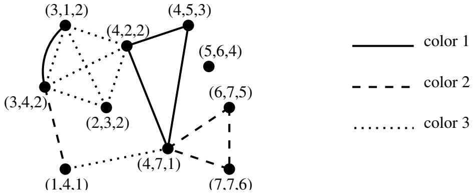


FIG. 1. *3-color graph representation.*

DEFINITION 1. *Let $\mathcal{G}$ be a 3-color graph. If, for some q-ary code $\mathcal{C}$ of length 3, we have $\mathcal{G} = \mathcal{C}^*$, then $\mathcal{G}$ is called a q-*ary code graph.

For a 3-color graph $\mathcal{G}$, for each $i = 1, 2, 3$, let $Color_i(\mathcal{G})$ denote the graph obtained from $\mathcal{G}$ by removing all the edges of colors other than color i.

It is easy to see that for each $i = 1, 2, 3$, $Color_i(\mathcal{C}^*)$ contains only cliques. This is also a sufficient condition that enables a 3-color graph to be a code graph, as stated in Theorem 2 below.

THEOREM 2. *If $\mathcal{G}$ is a q-ary code graph, then for all $i = 1, 2, 3$, $Color_i(\mathcal{G})$ contains only cliques of color i. Let $\#Color_i(\mathcal{G})$ denote the number of cliques in the graph $Color_i(\mathcal{G})$; then $\#Color_i(\mathcal{G}) \leq q$ for each $i = 1, 2, 3$.*

Conversely, for a 3-color graph $\mathcal{G}$, if for each $i = 1, 2, 3$ the graph $Color_i(\mathcal{G})$ only contains cliques and $\#Color_i(\mathcal{G}) \leq q$, then there exists a q-ary code $\mathcal{C}$ of length 3 such that $\mathcal{G} = \mathcal{C}^*$.

Given a q-ary code graph $\mathcal{G}$, we can construct a corresponding code in the following way. For each $i = 1, 2, 3$, label the cliques in the graph $Color_i(\mathcal{G})$ with the symbols 1, 2, ..., $\#Color_i(\mathcal{G})$. A node of $\mathcal{G}$ will be assigned a codeword (s_1, s_2, s_3) if it belongs to the clique of color 1 labeled s_1, the clique of color 2 labeled s_2, and the clique of color 3 labeled s_3 (see Figure 2). Thus, there are many codes associated with the

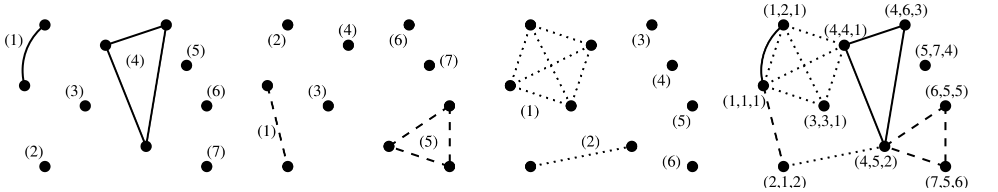


FIG. 2. *Constructing a code from the graph.*

same graph $\mathcal{G}$ that we consider *equivalent*. Equivalence of codes is as defined in [4].

DEFINITION 3 (see [4]). *Two q-ary codes $\mathcal{C}_1$, $\mathcal{C}_2$ of the same length n and same size m are equivalent if there exists a permutation σ of the n coordinate positions and permutations $\pi_1, \pi_2, \dots, \pi_n$ of the code alphabet for which $(c_1, c_2, \dots, c_n) \in \mathcal{C}_1$ if and only if $(\pi_1(c_{\sigma(1)}), \pi_2(c_{\sigma(2)}), \dots, \pi_n(c_{\sigma(n)})) \in \mathcal{C}_2$.*

3. 2-IPP *q*-ary code graph.

DEFINITION 4. *Let $\mathcal{G} = \mathcal{C}^*$ be a q-ary code graph for a code $\mathcal{C}$ of length 3. If the code $\mathcal{C}$ is 2-IPP, then $\mathcal{G}$ is called a* 2-IPP q-ary code graph *(*2-IPP code graph *for short).*

As stated in [3], it is easy to prove that $\mathcal{C}$ is a 2-IPP code if and only if the following conditions are satisfied:

(i) *IPP*1: a, b, c distinct in $\mathcal{C} \to a_i$, b_i, c_i distinct for some i.

(ii) *IPP*2: $a, b, c, d \in \mathcal{C}$ with $\{a, b\} \cap \{c, d\} = \emptyset \to \{a_i, b_i\} \cap \{c_i, d_i\} = \emptyset$ for some i.

Translating these conditions into graph language for $\mathcal{C}^*$ we have the following:

(i) *IPP*1: A, B, C distinct nodes in $\mathcal{C}^* \to$ there is a color i that does not appear in the edges joining these three nodes.

(ii) *IPP*2: For any two disjoint pairs of nodes $\{A, B\}$ and $\{C, D\}$ in $\mathcal{C}^*$, there is a color i that does not appear in the edges joining the two sets $\{A, B\}$ and $\{C, D\}$.

The patterns shown in Figure 3 are examples of forbidden patterns in 2-IPP graph.

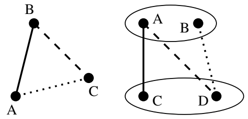


FIG. 3. *Forbidden patterns in 2-IPP graph.*

3.1. A partial ordering on the set of maximal code graphs.

Let $F(q) = F(3, q)$ denote the maximum size of a q-ary 2-IPP code of length 3. Our objective is to find $F(q)$. In [3], it is proved that for large enough q, $3q - 12\sqrt{q} \le F(q) \le 3q - 1$. There may exist many 2-IPP graphs with maximum number of nodes. We introduce a partial ordering on code graphs with a maximum number of nodes and find the maximum code graphs, which are also minimal based on this ordering.

First we introduce two orderings on the set of q-ary code graphs.

DEFINITION 5. *Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two q-ary code graphs. We say $\mathcal{G}_1$ is alphabetically smaller than $\mathcal{G}_2$, denoted by $\mathcal{G}_1 <_q \mathcal{G}_2$ if and only if*

(i) *$\#Color_i(\mathcal{G}_1) \le \#Color_i(\mathcal{G}_2)$ for all $i = 1, 2, 3$;*

(ii) *there is at least one i such that $\#Color_i(\mathcal{G}_1) < \#Color_i(\mathcal{G}_2)$.*

If $\#Color_i(\mathcal{G}_1) = \#Color_i(\mathcal{G}_2)$ for all $i = 1, 2, 3$, then $\mathcal{G}_1$ and $\mathcal{G}_2$ are said to be alphabetically equal, which is denoted by $\mathcal{G}_1 =_q \mathcal{G}_2$.

Intuitively, if $\mathcal{G}_1 <_q \mathcal{G}_2$, then the code generated by $\mathcal{G}_1$ uses less alphabet symbols than that of $\mathcal{G}_2$.

In general, we can consider $\mathcal{G}$ a union of its connected subgraphs. There are four types of connected subgraphs, which we will call 0-color part, 1-color part, 2-color part, and 3-color part, depending on the number of colors that appear in a particular subgraph. The 0-color part is just a single isolated node.

Since a 2-color part automatically satisfies the two conditions *IPP*1 and *IPP*2, it is preferable to a 3-color part.

DEFINITION 6. *Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two q-ary code graphs. We say $\mathcal{G}_1$ is more color* *efficient than $\mathcal{G}_2$ and denote it by $\mathcal{G}_1 \prec \mathcal{G}_2$, if and only if $\mathcal{G}_1$ has less* 3*-color parts than $\mathcal{G}_2$.*

Now we can define a partial order $\ll$ on the set of 2-IPP q-ary code graphs which have a maximum number of nodes.

Definition 7. *Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two* 2*-IPP q-ary code graphs with a maximum number of nodes ($|\mathcal{G}_1| = |\mathcal{G}_2| = F(q)$). Then $\mathcal{G}_1 \ll \mathcal{G}_2$ if and only if either $\mathcal{G}_1 <_q \mathcal{G}_2$ or, $\mathcal{G}_1 =_q \mathcal{G}_2$ and $\mathcal{G}_1 \prec \mathcal{G}_2$.*

From now on, we use $\mathcal{G}$ to denote a 2-IPP q-ary code graph which has a maximum number of nodes that is minimal in the ordering $\ll$. In section 4, we will establish properties of $\mathcal{G}$. Here we summarize all the known conditions on $\mathcal{G}$:

1. $\mathcal{G}$ is a 2-IPP q-ary code graph, which implies that,

(i) for each $i = 1, 2, 3$, $Color_i(\mathcal{G})$ consists of only cliques (*code graph condition*);

(ii) for each $i = 1, 2, 3$, $\#Color_i(\mathcal{G}) \leq q$ (*q-ary*);

(iii) for any three distinct nodes A, B and C, there exists a color i that does not appear among the edges that join these nodes (*IPP*1);

(iv) for any two disjoint pairs of nodes $\{A, B\}$ and $\{C, D\}$, there exists a color i that does not appear among the edges joining these two sets (*IPP*2).

2. If $\mathcal{G}'$ is a 2-IPP q-ary code graph, then $|\mathcal{G}'| \leq |\mathcal{G}| = F(q)$, where $|\mathcal{L}|$ denotes the number of nodes in the graph $\mathcal{L}$ (*maximal code*).

3. There does not exist a 2-IPP q-ary code graph $\mathcal{G}'$ with a maximum number of nodes $|\mathcal{G}'| = |\mathcal{G}|$ and $\mathcal{G}' <_q \mathcal{G}$ (*alphabetic minimal*).

4. There does not exist a 2-IPP q-ary code graph $\mathcal{G}'$ with a maximum number of nodes $|\mathcal{G}'| = |\mathcal{G}|$ that is alphabetically equal to $\mathcal{G}$, and $\mathcal{G}' \prec \mathcal{G}$ (*color efficient*).

4. Properties of $\mathcal{G}$.

It is easy to see that $F(1) = 1$ and $F(2) = 2$. From now on we assume $q > 2$.

Lemma 8. $|\mathcal{G}| = F(q) > q$ *(for all $q > 2$).*

Proof. A 2-IPP q-ary code graph with $q+1$ nodes is shown in Figure 4. Therefore, $|\mathcal{G}| \geq q + 1$. □

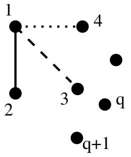


Fig. 4. *A* 2*-IPP q-ary code graph with $q + 1$ nodes.*

Lemma 9 (see Lemma 2 in [3]). *$\mathcal{G}$ is a simple graph. That is, any two nodes of $\mathcal{G}$ are joined by at most one edge.*

Proof. Assume that two nodes X and Y in $\mathcal{G}$ are joined by two edges of colors 1 and 2 (see Figure 5). We will show that in $\mathcal{G}$ the color 3 does not appear. Therefore, $|\mathcal{G}| = \#Color_3(\mathcal{G}) \leq q$, which contradicts Lemma 8.

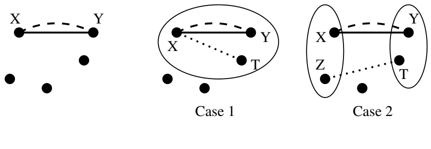


Fig. 5. *Nonsimple graph.*

Indeed, suppose we have edges of color 3. Then we have the following two cases.

Case 1. There is an edge of color 3 incident to X or Y. Without loss of generality, assume this edge joins a node T to a node X. Then three points X, Y, and T violate the condition *IPP*1.

Case 2. There is an edge of color 3 that joins a node T to a node Z. Then the pairs $\{Z, X\}$ and $\{T, Y\}$ violate the condition *IPP*2.

Therefore, $\mathcal{G}$ is a simple graph. $\square$

As noted in [3], the only possible 3-color part in $\mathcal{G}$ is a binding of three proper cliques of colors 1, 2, 3 at a single common node. Lemma 10 gives a complete proof of this statement, which is only outlined in [3]. We call a clique proper if its size > 1.

LEMMA 10. *A* 3*-color part* $\mathcal{P}$ *in* $\mathcal{G}$ *must be a binding of three proper cliques of colors* 1, 2, *and* 3 *at a common node (see Figure* 6*).*

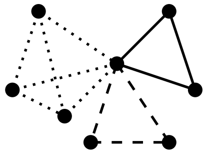


FIG. 6. *A* 3*-color part.*

Proof. Since $\mathcal{G}$ is simple, two proper cliques of different colors in $\mathcal{P}$, which have common nodes, must have exactly one common node; thus if we travel from one proper clique to another, we must pass through the common node.

First, we show that there is a node that belongs to three proper cliques of different colors. Suppose this is not the case. Then consider a path through proper cliques that passes each clique at most once. The sequence of common nodes in this path will consist of distinct nodes. Since $\mathcal{P}$ is a connected subgraph containing proper cliques of three colors, there exists a path that goes through three consecutive proper cliques of different colors (see Figure 7). So the pairs $\{A, C\}$ and $\{B, D\}$ violate the condition *IPP*2.

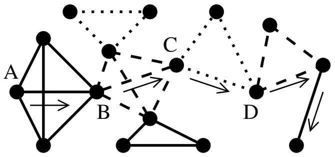


FIG. 7. *Travel through cliques.*

Therefore, there exists a common node B that belongs to three proper cliques $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ of colors 1, 2, 3 (see Figure 8). It is easy to see that $\mathcal{P}$ is a binding of only

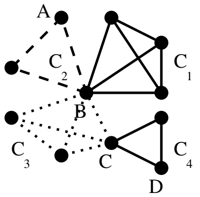


FIG. 8. $\mathcal{P}$ *contains more than three cliques.*

these three cliques $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ since, if $\mathcal{P}$ contains another clique, say $\mathcal{C}_4$, of color 1, then we again see the forbidden pattern $A \to B \to C \to D$. □

A 2-color part $\mathcal{P}$ of colors i and j can be represented as a rectangular grid which has $\#Color_i(\mathcal{P})$ rows and $\#Color_j(\mathcal{P})$ columns. Two nodes are in the same row if and only if they are in the same clique of color i; they are in the same column if and only if they are in the same clique of color j. We can also draw the grid such that the color i cliques are the columns and the color j cliques are the rows.

Note that a 2-color part contains at least three nodes because it has at least two rows and at least two columns.

Example. In Figure 9, the 2-color part $\mathcal{P}$ contains $\#Color_1(\mathcal{P}) = 4$ cliques of color 1 and $\#Color_2(\mathcal{P}) = 5$ cliques of color 2. It can be represented as a 4×5 rectangular grid, where the rows are color 1 cliques and the columns are color 2 cliques.

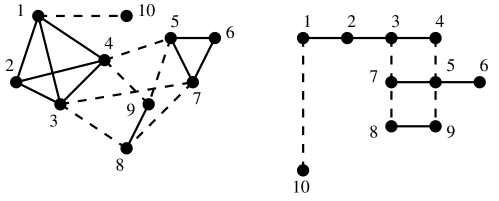


FIG. 9. *Rectangular grid representation of a* 2*-color part.*

LEMMA 11. *If $\mathcal{P}$ is a 2-color part in $\mathcal{G}$ and $\mathcal{P}$ contains n nodes, x_i cliques of color i, and x_j cliques of color j, then $\lceil n/x_i \rceil = x_j$ and $\lceil n/x_j \rceil = x_i$.*

Proof. Assume that $\lceil n/x_i \rceil = t < x_j$. Then we can construct a rectangular grid graph which has x_i rows and t columns containing n nodes. This grid is a new 2-color part $\mathcal{P}'$ which contains n nodes, x_i cliques of color i, and t cliques of color j (see Figure 10). The new graph $\mathcal{G}'$ obtained from $\mathcal{G}$ by replacing $\mathcal{P}$ with $\mathcal{P}'$ is also a 2-IPP q-ary code graph. It has the same number of nodes as $\mathcal{G}$ and $\mathcal{G}' <_q \mathcal{G}$, which is a contradiction. Therefore (because $n \leq x_i x_j$) we get $\lceil n/x_i \rceil = x_j$. Similarly, we have $\lceil n/x_j \rceil = x_i$. □

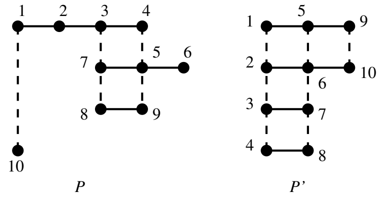
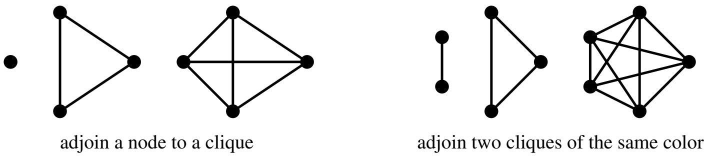


FIG. 10. *In $\mathcal{P}$: $n = 10$, $x_1 = 4$, $x_2 = 5$; in $\mathcal{P}'$: $n = 10$, $x_1 = 4$, $x_2 = 3$.*


adjoin a node to a clique

adjoin two cliques of the same color

FIG. 11. *Adjoining.*

In the next few lemmas, we will analyze the structure of $\mathcal{G}$. In the proofs, we often use the technique of adjoining a node to a clique or adjoining a clique to another clique of the same color. Here *adjoining* means joining all the nodes in question so that together they form a new clique of the same color (see Figure 11).

THEOREM 12. *$\mathcal{G}$ does not contain a 0-color part.*

Proof. Suppose that $\mathcal{G}$ contains a 0-color part, that is, an isolated node X. Since $|\mathcal{G}| > q$ (Lemma 8), not all connected subgraphs of $\mathcal{G}$ are 0-color parts. Let $\mathcal{P}$ be a connected subgraph of $\mathcal{G}$ which is not a 0-color part. Choose an arbitrary clique in $\mathcal{P}$ and adjoin X to that clique (see Figure 12). What we obtain is a new 2-IPP q-ary code graph which has the same number of nodes as $\mathcal{G}$ and is alphabetically smaller than $\mathcal{G}$. This is a contradiction. Therefore, $\mathcal{G}$ does not contain a 0-color part. □

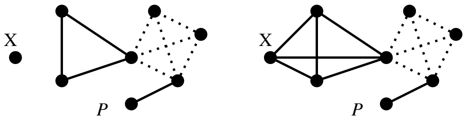


FIG. 12. *$\mathcal{G}$ contains a 0-color part.*

We will show in Lemmas 13, 14, and 15 that $\mathcal{G}$ can only contain at most one 1-color part, at most one 2-color part of the same colors, and at most one 3-color part. Two 2-color parts are of the same colors if the two pairs of colors appearing in them are the same.

LEMMA 13. *$\mathcal{G}$ does not contain more than one 1-color part.*

Proof. Suppose that $\mathcal{G}$ contains two 1-color parts.

Case 1. The two 1-color parts are of the same color. By adjoining the two parts, we have a new 2-IPP code graph which has the same number of nodes as $\mathcal{G}$ but is alphabetically smaller than $\mathcal{G}$ (a contradiction).

Case 2. The two 1-color parts are of different colors i and j. Replace the two 1-color parts with a new 2-color part by adjoining a node of the 1-color part of color i to the 1-color part of color j, as shown in Figure 13. We obtain a 2-IPP q-ary code graph which has the same number of nodes as $\mathcal{G}$ and is alphabetically smaller than $\mathcal{G}$ (a contradiction).

Therefore, $\mathcal{G}$ contains at most one 1-color part. □

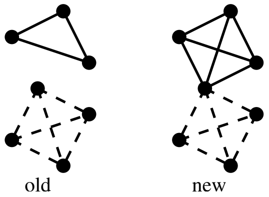


FIG. 13. *Case 2, two 1-color parts of different colors.*

LEMMA 14. *$\mathcal{G}$ does not contain two 2-color parts of the same colors.*

Proof. Suppose $\mathcal{G}$ contains two 2-color parts whose colors are the same. Select an arbitrary clique from one part, another clique of the same color from the other part, and adjoin them as shown in Figure 14. We obtain a new 2-IPP q-ary code graph

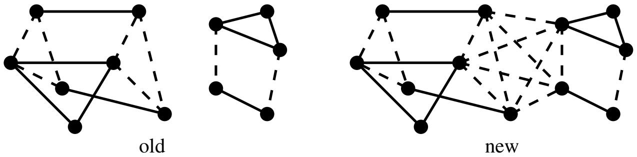


FIG. 14. *$\mathcal{G}$ contains two 2-color parts of colors 1 and 2.*

which has the same number of nodes as $\mathcal{G}$ and is alphabetically smaller than $\mathcal{G}$. This is a contradiction, and so $\mathcal{G}$ cannot contain two 2-color parts of the same colors. ∎

LEMMA 15. *$\mathcal{G}$ does not contain more than one 3-color part.*

Proof. Suppose that $\mathcal{G}$ contains two 3-color parts $\mathcal{P}_1$ and $\mathcal{P}_2$. $\mathcal{P}_1$ is a binding of a clique of color 1 of size x_1, a clique of color 2 of size y_1, and a clique of color 3 of size z_1. $\mathcal{P}_2$ is a binding of a clique of color 1 of size x_2, a clique of color 2 of size y_2, and a clique of color 3 of size z_2. Consider a new 3-color part $\mathcal{P}$ which is a binding of a clique of color 1 of size $x_1 + x_2 - 1$, a clique of color 2 of size $y_1 + y_2 - 1$, and a clique of color 3 of size $z_1 + z_2$ as shown in Figure 15.

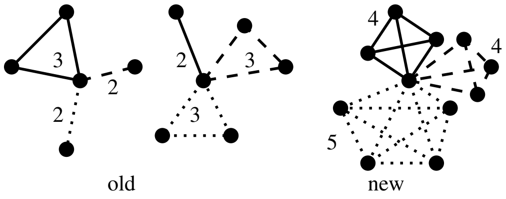


FIG. 15. *$\mathcal{G}$ contains two 3-color parts with $x_1 = 3$, $y_1 = 2$, $z_1 = 2$, $x_2 = 2$, $y_2 = 3$, $z_2 = 3$.*

We have

$$\begin{aligned}
|\mathcal{P}_1| &= x_1 + y_1 + z_1 - 2,\\
|\mathcal{P}_2| &= x_2 + y_2 + z_2 - 2,\\
|\mathcal{P}| &= (x_1 + x_2 - 1) + (y_1 + y_2 - 1) + (z_1 + z_2) - 2.
\end{aligned}$$

Therefore, $|\mathcal{P}| = |\mathcal{P}_1| + |\mathcal{P}_2|$.

We have

$$\begin{aligned}
\#Color_1(\mathcal{P}_1) &= 1 + (y_1 - 1) + (z_1 - 1),\\
\#Color_1(\mathcal{P}_2) &= 1 + (y_2 - 1) + (z_2 - 1),\\
\#Color_1(\mathcal{P}) &= 1 + (y_1 + y_2 - 1 - 1) + (z_1 + z_2 - 1).
\end{aligned}$$

Therefore, $\#Color_1(\mathcal{P}) = \#Color_1(\mathcal{P}_1) + \#Color_1(\mathcal{P}_2)$. Similarly, $\#Color_2(\mathcal{P}) = \#Color_2(\mathcal{P}_1) + \#Color_2(\mathcal{P}_2)$.

We have

$$\begin{aligned}
\#Color_3(\mathcal{P}_1) &= 1 + (x_1 - 1) + (y_1 - 1),\\
\#Color_3(\mathcal{P}_2) &= 1 + (x_2 - 1) + (y_2 - 1),\\
\#Color_3(\mathcal{P}) &= 1 + (x_1 + x_2 - 1 - 1) + (y_1 + y_2 - 1 - 1).
\end{aligned}$$

Therefore, $\#Color_3(\mathcal{P}) < \#Color_3(\mathcal{P}_1) + \#Color_3(\mathcal{P}_2)$.

So if we replace the two 3-color parts $\mathcal{P}_1$ and $\mathcal{P}_2$ with the new 3-color part $\mathcal{P}$, we obtain a new 2-IPP code graph that has the same number of nodes as $\mathcal{G}$ but is alphabetically smaller than $\mathcal{G}$, which is a contradiction. Therefore, $\mathcal{G}$ contains at most one 3-color part. ∎

LEMMA 16. *$\mathcal{G}$ does not contain both a 1-color part and a 3-color part.*

Proof. Suppose that $\mathcal{G}$ contains a 1-color part, that is, a clique, and a 3-color part. Adjoin the 1-color part to the clique of the same color in the 3-color part as shown in Figure 16. We obtain a new 2-IPP code graph which has the same number of nodes as $\mathcal{G}$ and is alphabetically smaller than $\mathcal{G}$. This is a contradiction. Therefore, $\mathcal{G}$ cannot contain a 1-color part and a 3-color part at the same time. ∎

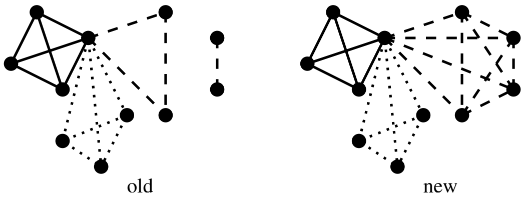


FIG. 16. *$\mathcal{G}$ contains both* 1*-color part and* 3*-color part.*

LEMMA 17. *$\mathcal{G}$ does not contain a* 2*-color part and a* 1*-color part of the color which appears in the* 2*-color part.*

Proof. Suppose that $\mathcal{G}$ contains a 1-color part of color i and a 2-color part made of color i and another color. Select an arbitrary clique of color i in the 2-color part and adjoin it to the 1-color part (see Figure 17). Then we have a new 2-IPP q-ary code graph that has the same number of nodes as $\mathcal{G}$ and is alphabetically smaller than $\mathcal{G}$. This is a contradiction, and so $\mathcal{G}$ cannot have a 2-color part and a 1-color part of a color that appears in the 2-color part. ▯

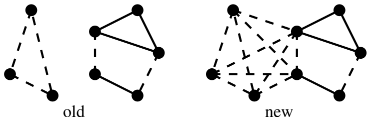


FIG. 17. *$\mathcal{G}$ contains a* 1*-color part of color* 2 *and a* 2*-color part of colors* 1 *and* 2*.*

LEMMA 18. *If $\mathcal{G}$ contains a* 3*-color part $\mathcal{P}$ (which is a binding of three proper cliques of colors* 1, 2, 3*), then at least two of the three cliques in $\mathcal{P}$ must be of size* 2*.*

Proof. For each $i = 1, 2, 3$, let k_i be the size of the proper clique of color i in $\mathcal{P}$. Without loss of generality, we may assume that $k_1 \geq k_2 \geq k_3$. We need to show that $k_2 = k_3 = 2$. Suppose this is not the case; then we have $k_2 \geq 3$.

Construct a 1-color part $\mathcal{P}_1$ of color 1 and size k_3. Let $\mathcal{P}_2$ be a 2-color part whose rectangular grid representation has two rows corresponding to two cliques of color 1. The first row contains $k_1 - 1$ nodes and the second contains $k_2 - 1$ nodes. There are $k_1 - 1$ columns. Each of the first $k_2 - 1$ columns contains two nodes and forms $k_2 - 1$ cliques of color 2, each of size 2 (see Figure 18).

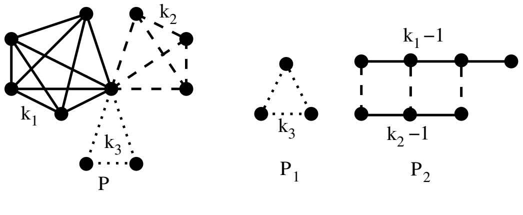


FIG. 18. *Breaking the* 3*-color part to form a better graph.*

We have $|\mathcal{P}| = k_1 + k_2 + k_3 - 2$, $|\mathcal{P}_1| = k_3$, and $|\mathcal{P}_2| = (k_1 - 1) + (k_2 - 1)$. Therefore, $|\mathcal{P}| = |\mathcal{P}_1| + |\mathcal{P}_2|$.

We have $\#Color_1(\mathcal{P}) = 1+(k_2-1)+(k_3-1)$, $\#Color_1(\mathcal{P}_1) = k_3$, and $\#Color_1(\mathcal{P}_2) = 2$. Since $k_2 \geq 3$, we have $\#Color_1(\mathcal{P}) \geq \#Color_1(\mathcal{P}_1) + \#Color_1(\mathcal{P}_2)$.

Now

$$\#Color_2(\mathcal{P}) = 1 + (k_3 - 1) + (k_1 - 1),$$
$$\#Color_2(\mathcal{P}_1) = k_3, \#Color_2(\mathcal{P}_2) = k_1 - 1,$$

imply that $\#Color_2(\mathcal{P}) = \#Color_2(\mathcal{P}_1) + \#Color_2(\mathcal{P}_2)$. Moreover,

$$\#Color_3(\mathcal{P}) = 1 + (k_1 - 1) + (k_2 - 1),$$
$$\#Color_3(\mathcal{P}_1) = 1,$$
$$\#Color_3(\mathcal{P}_2) = (k_1 - 1) + (k_2 - 1)$$

imply $\#Color_3(\mathcal{P}) = \#Color_3(\mathcal{P}_1) + \#Color_3(\mathcal{P}_2)$.

Therefore, if we replace $\mathcal{P}$ with $\mathcal{P}_1$ and $\mathcal{P}_2$, then we have a new 2-IPP code graph $\mathcal{G}'$ which has the same number of nodes as $\mathcal{G}$ and $\mathcal{G}' \ll \mathcal{G}$. This is a contradiction.

So $k_2 = 2$, which implies that $k_2 = k_3 = 2$, and the 3-color part $\mathcal{P}$ in $\mathcal{G}$ must contain at least two proper cliques of size 2 (see Figure 19). □

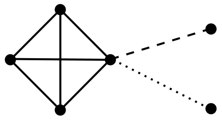


FIG. 19. *Shape of the 3-color part in $\mathcal{G}$.*

5. Graph structure of $\mathcal{G}$. Based on the properties of $\mathcal{G}$ proved in the previous section, we can now determine the structure of $\mathcal{G}$. Theorems 19, 20, and 21 below show that $\mathcal{G}$ can be one of the four possible types.

THEOREM 19. *If $\mathcal{G}$ does not contain a 1-color part or a 3-color part, then $\mathcal{G}$ consists of either two or three 2-color parts of different color types.*

If $\mathcal{G}$ consists of three 2-color parts, then $q \geq 7$ and we say that $\mathcal{G}$ is of type I.

If $\mathcal{G}$ consists of two 2-color parts, then $q \geq 5$ and $|\mathcal{G}| = 2q - 4$. In this case, we say that $\mathcal{G}$ is of type II.

Proof. If $\mathcal{G}$ does not contain a 1-color part or a 3-color part, then $\mathcal{G}$ contains only 2-color parts. Lemma 14 says that $\mathcal{G}$ cannot contain two 2-color parts of the same color; therefore $\mathcal{G}$ is a union of at most three 2-color parts.

$\mathcal{G}$ contains at least two parts since, if it consists of only one 2-color part, say of color 1 and 2, then $|\mathcal{G}| = \#Color_3(\mathcal{G}) \leq q$, which contradicts Lemma 8. So $\mathcal{G}$ consists of either two 2-color parts or three 2-color parts.

Case 1. Let $\mathcal{G}$ contain three 2-color parts. Use n_1 to denote the number of nodes in the 2-color part of colors 2 and 3 and use y_1, z_1 to denote the number of cliques of color 2 and 3, respectively, in this part; use n_2 to denote the number of nodes in the part of colors 3 and 1, and use z_2, x_2 to denote the number of cliques of color 3 and 1, respectively, in this part; finally, use n_3 to denote the number of nodes in the part of colors 1 and 2, and use x_3, y_3 to denote the number of cliques of color 1 and 2, respectively, in this part (see Figure 20). Then $\#Color_1(\mathcal{G}) = n_1 + x_2 + x_3 \leq q \rightarrow q \geq 3 + 2 + 2 = 7$.

Case 2. If $\mathcal{G}$ contains two 2-color parts (see Figure 21), then,

$$\#Color_1(\mathcal{G}) = n_1 + x_2 \leq q, \#Color_2(\mathcal{G}) = n_2 + y_1 \leq q, \#Color_3(\mathcal{G}) = z_1 + z_2 \leq q.$$

This implies that $q \geq 5$ and $|\mathcal{G}| = n_1 + n_2 \leq 2q - x_2 - y_1 \leq 2q - 4$.

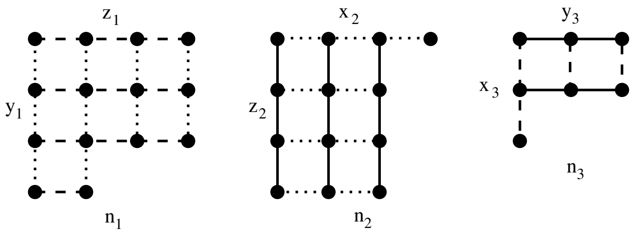


FIG. 20. *Graph of type* I.

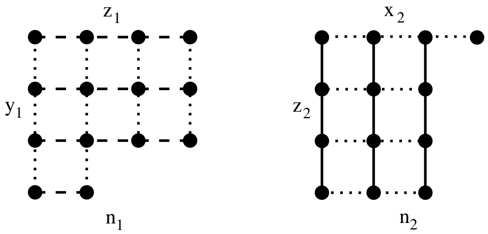


FIG. 21. *Graph of type* II.

In fact we can make $|\mathcal{G}| = 2q - 4$ by choosing $y_1 = x_2 = 2$, $n_1 = n_2 = q - 2$, $z_1 = z_2 = \lceil (q-2)/2 \rceil$. □

THEOREM 20. *If $\mathcal{G}$ contains a* 1*-color part, then $\mathcal{G}$ has exactly two connected subgraphs: the* 1*-color part and a* 2*-color part of the colors different from the color of the* 1*-color part.*

In this case, we have $q \geq 4$ and $|\mathcal{G}| \leq 2q - 3$. We call $\mathcal{G}$ type III*. Furthermore, if $q \geq 6$, then $|\mathcal{G}| \leq 2q - 4$.*

Proof. If $\mathcal{G}$ contains a 1-color part $\mathcal{P}_1$, then Lemma 13 says that this is the unique 1-color part; Lemma 16 excludes the existence of a 3-color part. Since $|\mathcal{G}| > q$ (Lemma 8), $\mathcal{G}$ must contain connected subgraphs other than the 1-color part. The only possible form of these subgraphs is to be 2-color parts. Let us assume $\mathcal{P}_1$ is of color 1. Then Lemma 17 says that 2-color parts in $\mathcal{G}$ must be made of colors 2 and 3. Lemma 14 says that there is only one such 2-color part in $\mathcal{G}$—call it $\mathcal{P}_2$.

Thus, $\mathcal{G}$ is the union of the 1-color part $\mathcal{P}_1$ of color 1 and the 2-color part $\mathcal{P}_2$ of colors 2 and 3 (see Figure 22).

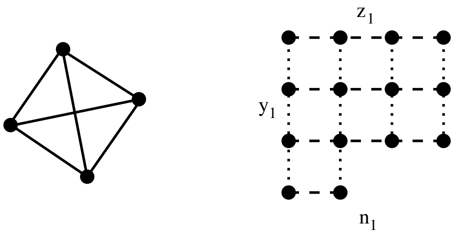


FIG. 22. *Graph of type* III.

We have $\#Color_1(\mathcal{G}) = 1 + n_1 \leq q$, hence $q \geq 4$ and $n_1 \leq q - 1$. We have

$$\#Color_2(\mathcal{G}) = |\mathcal{P}_1| + y_1 \leq q \rightarrow |\mathcal{P}_1| \leq q - y_1,$$
$$\#Color_3(\mathcal{G}) = |\mathcal{P}_1| + z_1 \leq q \rightarrow |\mathcal{P}_1| \leq q - z_1.$$

We have $y_1, z_1 \geq 2$, and thus $|\mathcal{P}_1| \leq q-2$ and $|\mathcal{G}| = |\mathcal{P}_1| + n_1 \leq 2q-3$.

If $|\mathcal{G}| = 2q-3$, then $y_1 = z_1 = 2$ and $n_1 = q-1 \leq y_1 z_1 = 4 \to q \leq 5$. Therefore if $q \geq 6$, then $|\mathcal{G}| \leq 2q-4$. □

THEOREM 21. *If $\mathcal{G}$ contains a* 3*-color part $\mathcal{P}$, then $\mathcal{G} = \mathcal{P}$, and $\mathcal{G}$ is a binding of three cliques of sizes* 2, 2*, and* $q-1$*. In this case, we say $\mathcal{G}$ is type* IV *and we have* $|\mathcal{G}| = q+1$.

Proof. From Lemma 18, we can assume that $\mathcal{P}$ is a binding of a clique of color 1 of size k and two cliques of colors 2 and 3, both of size 2. We need to show that $\mathcal{P}$ is the whole of $\mathcal{G}$ and $k = q-1$.

Let A and B, respectively, be two nodes in the clique of color 2 and the clique of color 3 of $\mathcal{P}$ such that neither A nor B is the 3-color part's common node.

Suppose $\mathcal{P}$ is not the whole $\mathcal{G}$. Then $\mathcal{G}$ has another connected subgraph $\mathcal{Q}$. Theorem 12, Lemma 15, and Lemma 16 ensure that $\mathcal{Q}$ is a 2-color part. There are three cases.

Case 1. $\mathcal{Q}$ is made up of colors 2 and 3 (Figure 23).

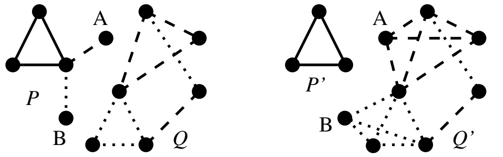


FIG. 23. *$\mathcal{G}$ contains a* 3*-color part and a* 2*-color part.*

The 1-color part $\mathcal{P}'$ is obtained from $\mathcal{P}$ by disconnecting A and B from $\mathcal{P}$ and adjoining A and B to arbitrary cliques of color 2 and 3, respectively, in $\mathcal{Q}$ to get the 2-color part $\mathcal{Q}'$.

Clearly, $|\mathcal{P}'| + |\mathcal{Q}'| = |\mathcal{P}| + |\mathcal{Q}|$, and

$$\#Color_1(\mathcal{P}') + \#Color_1(\mathcal{Q}') = \#Color_1(\mathcal{P}) + \#Color_1(\mathcal{Q}).$$

Moreover,

$$\begin{aligned} &\#Color_2(\mathcal{P}') = k, \\ &\#Color_2(\mathcal{Q}') = \#Color_2(\mathcal{Q}) + 1, \\ &\#Color_2(\mathcal{P}) = k+1 \end{aligned}$$

imply that $\#Color_2(\mathcal{P}') + \#Color_2(\mathcal{Q}') = \#Color_2(\mathcal{P}) + \#Color_2(\mathcal{Q})$.

Similarly, $\#Color_3(\mathcal{P}') + \#Color_3(\mathcal{Q}') = \#Color_3(\mathcal{P}) + \#Color_3(\mathcal{Q})$.

Therefore, if we replace $\mathcal{P}$ and $\mathcal{Q}$ with $\mathcal{P}'$ and $\mathcal{Q}'$, then we have a new 2-IPP q-ary code graph $\mathcal{G}'$ with the same number of nodes as $\mathcal{G}$ and $\mathcal{G}' \ll \mathcal{G}$, which is a contradiction. *Case* 2. $\mathcal{Q}$ is made up of colors 3 and 1 (Figure 24).

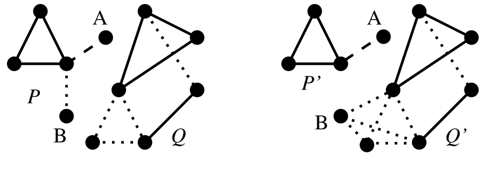


FIG. 24. *$\mathcal{G}$ contains a* 3*-color part and a* 2*-color part.*

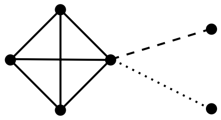


FIG. 25. *$\mathcal{G}$ is of type* IV.

Disconnect B from part $\mathcal{P}$ and adjoin it to an arbitrary clique of color 3 in $\mathcal{Q}$ to obtain $\mathcal{P}'$ and $\mathcal{Q}'$.

Clearly, $|\mathcal{P}'| + |\mathcal{Q}'| = |\mathcal{P}| + |\mathcal{Q}|$,

$$\#Color_1(\mathcal{P}') + \#Color_1(\mathcal{Q}') = \#Color_1(\mathcal{P}) + \#Color_1(\mathcal{Q}), \text{ and}$$
$$\#Color_2(\mathcal{P}') + \#Color_2(\mathcal{Q}') = \#Color_2(\mathcal{P}) + \#Color_2(\mathcal{Q}).$$

Moreover, $\#Color_3(\mathcal{P}') = \#Color_3(\mathcal{P}) = k+1$, $\#Color_3(\mathcal{Q}') = \#Color_3(\mathcal{Q})$ imply that $\#Color_3(\mathcal{P}') + \#Color_3(\mathcal{Q}') = \#Color_3(\mathcal{P}) + \#Color_3(\mathcal{Q})$.

Therefore, if we replace $\mathcal{P}$ and $\mathcal{Q}$ with $\mathcal{P}'$ and $\mathcal{Q}'$, then we have a new 2-IPP code graph $\mathcal{G}'$ which has the same number of nodes as $\mathcal{G}$ and $\mathcal{G}' \ll \mathcal{G}$—a contradiction.

Case 3. $\mathcal{Q}$ is made up of colors 1 and 2. The proof is similar to that of Case 2.

Therefore, $\mathcal{P}$ is the whole $\mathcal{G}$.

We have $|\mathcal{G}| = k+2$ and $\#Color_2(\mathcal{G}) = k+1 \leq q$ (see Figure 25). Hence, $|\mathcal{G}| \leq q+1$ but, according to Lemma 8, $|\mathcal{G}| > q$ and so $|\mathcal{G}| = q+1$, which implies $k = q-1$. □

COROLLARY 22. *For $q \geq 5$, $|\mathcal{G}| \geq q+2$; hence $\mathcal{G}$ is not of type* IV.

Proof. For $q \geq 5$, there exists a 2-IPP code graph with $q+2$ nodes (see Figure 26). Therefore $|\mathcal{G}| \geq q+2$. □

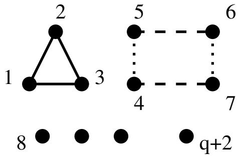


FIG. 26. *A 2-IPP q-ary graph with $q+2$ nodes.*

THEOREM 23. *For $q \geq 17$, $\mathcal{G}$ consists of three 2-color parts of different color types. This means that $\mathcal{G}$ is of type* I.

Proof. In section 6, we show the construction of a 2-IPP code graph consisting of three 2-color parts with more than $2q-4$ nodes if $q \geq 17$ (Corollary 25). Using this construction and the bounds derived on the sizes of type II, III, and IV code graphs, we conclude that $\mathcal{G}$ cannot be of type II, III, or IV; it must be of type I. □

6. The main theorem. In this section, for each $q \geq 8$ we will construct a 2-IPP code graph $\mathcal{G}_0$ and prove our main result, that is, $|\mathcal{G}_0| \leq F(q) \leq |\mathcal{G}_0| + 2$. As in previous sections, $\mathcal{G}$ denotes a maximal 2-IPP q-ary code graph which is minimal in the ordering $\ll$.

Construction of the graph $\mathcal{G}_0$. There are six cases:

Case 1. $q = r^2 + 2r$, where $r \geq 2$ (Figure 27). We have

$$\#Color_1(\mathcal{G}_0) = \#Color_2(\mathcal{G}_0) = \#Color_3(\mathcal{G}_0) = q,$$
$$|\mathcal{G}_0| = 3r^2 = 3q + 6 - 6(r+1) = 3q + 6 - 6\sqrt{q+1} = 3q + 6 - 6\lceil \sqrt{q+1}\, \rceil.$$

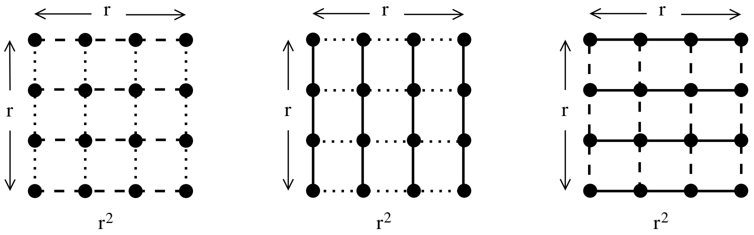


FIG. 27. *Graph $\mathcal{G}_0$ when $q = r^2 + 2r$.*

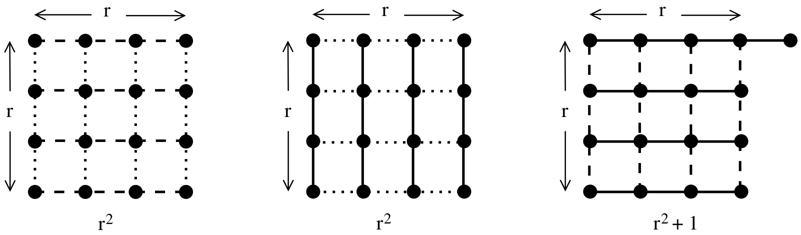


FIG. 28. *Graph $\mathcal{G}_0$ when $q = r^2 + 2r + 1$.*

Note that, in this case, the code generated by $\mathcal{G}_0$ is equivalent to the code given in Example 4 in [3].

Case 2. $q = r^2 + 2r + 1$, where $r \geq 2$ (Figure 28). We have

$$\#Color_1(\mathcal{G}_0) = q - 1, \#Color_2(\mathcal{G}_0) = \#Color_3(\mathcal{G}_0) = q,$$
$$|\mathcal{G}_0| = 3r^2 + 1 = 3q + 10 - 6(r+2) = 3q + 10 - 6\lceil \sqrt{q+1}\ \rceil.$$

Case 3. $q = r^2 + 2r + 2$, where $r \geq 2$ (Figure 29). We have

$$\#Color_1(\mathcal{G}_0) = \#Color_2(\mathcal{G}_0) = \#Color_3(\mathcal{G}_0) = q,$$
$$|\mathcal{G}_0| = 3r^2 + 4 = 3q + 10 - 6(r+2) = 3q + 10 - 6\lceil \sqrt{q+1}\ \rceil.$$

Case 4. $q = r^2 + 2r + 1 + k$, where $2 \leq k \leq r$ (Figure 30). We have

$$\#Color_1(\mathcal{G}_0) = \#Color_2(\mathcal{G}_0) = \#Color_3(\mathcal{G}_0) = q,$$
$$|\mathcal{G}_0| = 3(r^2 + k) = 3q + 9 - 6(r+2) = 3q + 9 - 6\lceil \sqrt{q+1}\ \rceil.$$

Case 5. $q = r^2 + 3r + 2$, where $r \geq 2$ (Figure 31). We have

$$\#Color_1(\mathcal{G}_0) = \#Color_2(\mathcal{G}_0) = \#Color_3(\mathcal{G}_0) = q,$$
$$|\mathcal{G}_0| = 3r^2 + 3r + 2 = 3q + 8 - 6(r+2) = 3q + 8 - 6\lceil \sqrt{q+1}\ \rceil.$$

Case 6. $q = r^2 + 3r + 2 + k$, where $r \geq 2$ and $1 \leq k \leq r$ (Figure 32). We have

$$\#Color_1(\mathcal{G}_0) = \#Color_2(\mathcal{G}_0) = \#Color_3(\mathcal{G}_0) = q,$$
$$|\mathcal{G}_0| = 3(r^2 + r + k) = 3q + 6 - 6(r+2) = 3q + 6 - 6\lceil \sqrt{q+1}\ \rceil.$$

COROLLARY 24. *For all $q \geq 8$, $F(q) \geq |\mathcal{G}_0| \geq 3q + 6 - 6\lceil \sqrt{q+1}\ \rceil$.*

Proof. The proof follows from the above construction and from noting that $r \geq 2$ when $q \geq 8$. ◻

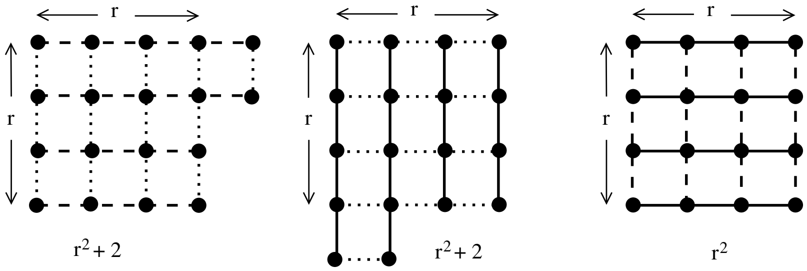


FIG. 29. *Graph $\mathcal{G}_0$ when $q = r^2 + 2r + 2$.*

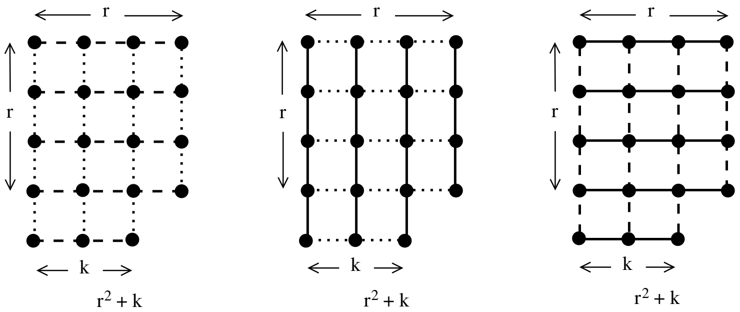


FIG. 30. *Graph $\mathcal{G}_0$ when $q = r^2 + 2r + 1 + k$, where $2 \le k \le r$.*

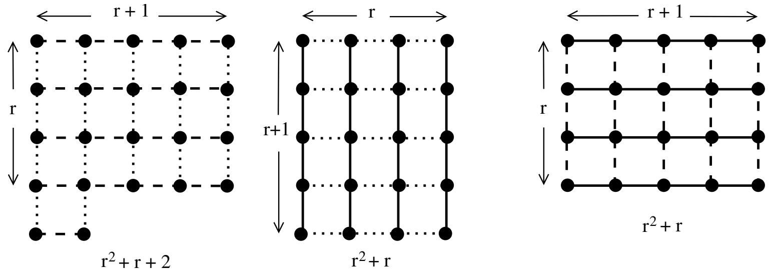


FIG. 31. *Graph $\mathcal{G}_0$ when $q = r^2 + 3r + 2$.*

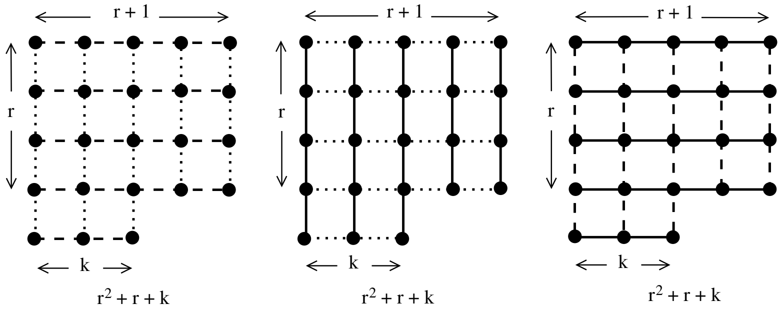


FIG. 32. *Graph $\mathcal{G}_0$ when $q = r^2 + 3r + 2 + k$, where $1 \le k \le r$.*

COROLLARY 25. *$|\mathcal{G}_0| > 2q - 4$ for all $q \ge 17$.*
Proof. Consider the following six cases.

Case 1. $q = r^2 + 2r$. $q \geq 17 \to r > 3$. We have

$$|\mathcal{G}_0| = 3r^2 \to |\mathcal{G}_0| - (2q-4) = 3r^2 - (2r^2 + 4r - 4) = r^2 - 4r + 4 = (r-2)^2 > 0.$$

Case 2. $q = r^2 + 2r + 1$. $q \geq 17 \to r > 3$. We have

$$|\mathcal{G}_0| = 3r^2 + 1 \to |\mathcal{G}_0| - (2q-4) = 3r^2 + 1 - (2r^2 + 4r - 2) = (r-2)^2 - 1 > 0.$$

Case 3. $q = r^2 + 2r + 2$. $q \geq 17 \to r \geq 3$. We have

$$|\mathcal{G}_0| = 3r^2 + 4 \to |\mathcal{G}_0| - (2q-4) = 3r^2 + 4 - (2r^2 + 4r) = r^2 - 4r + 4 = (r-2)^2 > 0.$$

Case 4. $q = r^2 + 2r + 1 + k$, where $2 \leq k \leq r$. $q \geq 17 \to r > 2$. We have

$$|\mathcal{G}_0| = 3(r^2+k) \to |\mathcal{G}_0| - (2q-4) = 3(r^2+k) - (2r^2+4r+2k-2) = (r-2)^2 + k - 2 > 0.$$

Case 5. $q = r^2 + 3r + 2$. $q \geq 17 \to r > 2$. We have

$$|\mathcal{G}_0| = 3r^2 + 3r + 2 \to |\mathcal{G}_0| - (2q-4) = 3r^2 + 3r + 2 - (2r^2 + 6r) = (r-2)(r-1) > 0.$$

Case 6. $q = r^2 + 3r + 2 + k$, where $1 \leq k \leq r$. $q \geq 17 \to r > 2$. We have

$$\begin{aligned}|\mathcal{G}_0| = 3(r^2 + r + k) \to |\mathcal{G}_0| - (2q-4) &= 3(r^2 + r + k) - (2r^2 + 6r + 2k) \\ &= (r-2)(r-1) + k - 2 > 0.\end{aligned}$$

Therefore, $|\mathcal{G}_0| > 2q - 4$ for all $q \geq 17$. □

DEFINITION 26. *Let q be an integer, $q \geq 5$. Let $n(q)$ denote the maximum value of $n_1 + n_2 + n_3$, where n_1, n_2, n_3, x_2, x_3, y_3, y_1, z_1, z_2 are positive integers satisfying the following conditions:*

$$n_1 \leq y_1 z_1,\ n_2 \leq z_2 x_2,\ n_3 \leq x_3 y_3,$$
$$n_1 + x_2 + x_3 \leq q,\ n_2 + y_3 + y_1 \leq q,\ n_3 + z_1 + z_2 \leq q,$$
$$x_2,\ x_3,\ y_3,\ y_1,\ z_1,\ z_2 \geq 2.$$

The following theorem is proved in the appendix.

THEOREM 27. *For all $q \geq 5$, $n(q) \leq 3q + 6 - 6\sqrt{q+1}$.*

LEMMA 28. *For all $q \geq 5$, $F(q) \geq n(q)$.*

Proof. We prove $F(q) \geq n(q)$ by showing that if positive integers n_1, n_2, n_3, x_2, x_3, y_3, y_1, z_1, z_2 satisfy all the conditions stated in Definition 26, then there exists a 2-IPP q-ary code graph with $n_1 + n_2 + n_3$ nodes.

Indeed, since $n_1 \leq y_1 z_1$, it is possible to construct a graph P_1 with n_1 nodes such that $\#Color_1(P_1) = n_1$, $\#Color_2(P_1) \leq y_1$, and $\#Color_3(P_1) \leq z_1$. If $1 \leq n_1 \leq 2$, then P_1 can be a 1-color part (of color 2 or 3) and; if $n_1 \geq 3$, then P_1 can be a 2-color part of colors 2 and 3 whose rectangular grid has up to y_1 rows and z_1 columns. Since $n_2 \leq z_2 x_2$ and $n_3 \leq x_3 y_3$, it is possible to construct similar graphs P_2 and P_3 with n_2 and n_3 nodes, respectively, and $\#Color_2(P_2) = n_2$, $\#Color_3(P_2) \leq z_2$, $\#Color_1(P_2) \leq x_2$, $\#Color_3(P_3) = n_3$, $\#Color_1(P_3) \leq x_3$, $\#Color_2(P_3) \leq y_3$. Let P be the union of P_1, P_2, and P_3; then P is a graph with $n_1 + n_2 + n_3$ nodes and does not contain any 3-color part. Moreover,

$$\#Color_1(P) = \#Color_1(P_1) + \#Color_1(P_2) + \#Color_1(P_3) \leq n_1 + x_2 + x_3 \leq q,$$
$$\#Color_2(P) = \#Color_2(P_1) + \#Color_2(P_2) + \#Color_2(P_3) \leq y_1 + n_2 + y_3 \leq q,$$
$$\#Color_3(P) = \#Color_3(P_1) + \#Color_3(P_2) + \#Color_3(P_3) \leq z_1 + z_2 + n_3 \leq q.$$

Therefore, P is a 2-IPP q-ary code graph with $n_1 + n_2 + n_3$ nodes. □

LEMMA 29. *For all $q \geq 17$, $F(q) \leq n(q)$.*

Proof. From Theorem 23 we know that, for $q \geq 17$, graph $\mathcal{G}$ must contain exactly three 2-color parts of different colors (Figure 33). This means that n_1, n_2, n_3, x_2, x_3, y_3, y_1, z_1, z_2 are positive integers satisfying all the conditions stated in Definition 26. Therefore, $|\mathcal{G}| = F(q) = n_1 + n_2 + n_3 \leq n(q)$. □

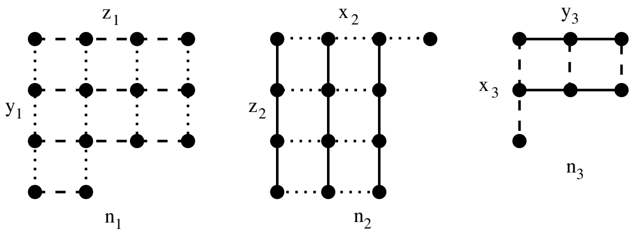


FIG. 33. *$\mathcal{G}$ must be of type* I *when $q \geq 17$.*

THEOREM 30. *For all $q \geq 17$, $F(q) = n(q)$.*

Proof. The proof follows from Lemmas 28 and 29. □

LEMMA 31. *For all $q \geq 17$,* $\quad F(q) \leq 3q + 6 - \lceil 6\sqrt{q+1}\,\rceil \leq |\mathcal{G}_0| + 2$.

Proof. From Theorems 30 and 27, for all $q \geq 17$ we have $F(q) = n(q) \leq [3q + 6 - 6\sqrt{q+1}] = 3q + 6 - \lceil 6\sqrt{q+1}\,\rceil$.

Let $f(q) = 3q + 6 - 6\sqrt{q+1}$. We complete the proof by showing that $[f(q)] \leq |\mathcal{G}_0| + 2$. Again, consider six cases.

Case 1. $q = r^2 + 2r$,

$$|\mathcal{G}_0| = 3q + 6 - 6\sqrt{q+1} = f(q) = [f(q)].$$

Case 2. $q = r^2 + 2r + 1$,

$$\sqrt{q+1} > r + 1 \to f(q) = 3q + 6 - 6\sqrt{q+1} < 3q + 6 - 6(r+1) = 3q - 6r$$
$$\to [f(q)] \leq 3q - 6r - 1 = 3(r^2 + 2r + 1) - 6r - 1 = 3r^2 + 2 = |\mathcal{G}_0| + 1.$$

Case 3. $q = r^2 + 2r + 2$. Similarly to Case 2, we have $[f(q)] \leq 3q - 6r - 1 = 3r^2 + 5 = |\mathcal{G}_0| + 1$.

Case 4. $q = r^2 + 2r + 1 + k$, where $2 \leq k \leq r$. Similarly to Case 2, we have $[f(q)] \leq 3q - 6r - 1 = 3r^2 + 3k + 2 = |\mathcal{G}_0| + 2$.

Case 5. $q = r^2 + 3r + 2$,

$$\sqrt{q+1} > r + 3/2 \to f(q) = 3q + 6 - 6\sqrt{q+1} < 3q + 6 - 6(r + 3/2) = 3q - 6r - 3$$
$$\to [f(q)] \leq 3q - 6r - 4 = 3(r^2 + 3r + 2) - 6r - 4 = 3r^2 + 3r + 2 = |\mathcal{G}_0|.$$

Case 6. $q = r^2 + 3r + 2 + k$, where $1 \leq k \leq r$. Similarly to Case 5, we have $[f(q)] \leq 3q - 6r - 4 = 3r^2 + 3r + 3k + 2 = |\mathcal{G}_0| + 2$.

In all six cases, we have $[f(q)] \leq |\mathcal{G}_0| + 2$. □

The above proofs of Lemmas 29 and 31 also show the following.

COROLLARY 32. *If the graph $\mathcal{G}$ is of type* I*, then*

$$|\mathcal{G}| \leq 3q + 6 - \lceil 6\sqrt{q+1}\,\rceil.$$

In particular, when q is of form $r^2 + 2r$ or $r^2 + 3r + 2$, for $r \geq 2$, and graph $\mathcal{G}$ is of type I*, then*

$$|\mathcal{G}| = 3q + 6 - \lceil 6\sqrt{q+1}\,\rceil = |\mathcal{G}_0|.$$

Combining the results from Corollary 24, Lemma 31, and Corollary 32, we have the following main theorem.

THEOREM 33. *For all $q \geq 17$,*

$$3q + 6 - 6\lceil \sqrt{q+1}\, \rceil \leq |\mathcal{G}_0| \leq F(q) \leq 3q + 6 - \lceil 6\sqrt{q+1}\, \rceil \leq |\mathcal{G}_0| + 2.$$

Particularly, if $q = r^2 + 2r$ or $q = r^2 + 3r + 2$, then $F(q) = |\mathcal{G}_0|$.

THEOREM 34. *For $3 \leq q \leq 16$ we have*

q	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$F(q)$	4	5	7	8	10	12	14	16	18	20	22	24	27	28

.

Proof. $q = 3$: Theorems 19 and 20 imply that $\mathcal{G}$ cannot be of type I, II, or III. So $\mathcal{G}$ is of type IV and, from Theorem 21, we have $|\mathcal{G}| = q + 1 = 4$, and $\mathcal{G}$ is a binding of three cliques of sizes 2, 2, and $q - 1 = 2$ (see Figure 34). Therefore $F(3) = 4$.

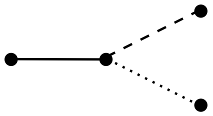


FIG. 34. *A maximal graph when $q = 3$.*

$q = 4$: Theorem 19 implies that $\mathcal{G}$ cannot be of type I or II. So $\mathcal{G}$ is of type III or IV. If it is of type III, then by Theorem 20, $|\mathcal{G}| \leq 2q - 3 = 5$. If it is of type IV, then by Theorem 21, $|\mathcal{G}| = q + 1 = 5$, and $\mathcal{G}$ is a binding of three cliques of sizes 2, 2, and $q - 1 = 3$ (see Figure 35). Therefore $F(4) = 5$.

$q = 5$: Theorem 19 and Corollary 22 imply that $\mathcal{G}$ cannot be of type I or IV. So $\mathcal{G}$ is of type II or III. If it is of type II, then by Theorem 19, $|\mathcal{G}| = 2q - 4 = 6$. If it is of type III, then by Theorem 20, $|\mathcal{G}| \leq 2q - 3 = 7$. Figure 36 shows a graph of type III with 7 nodes, therefore implying $F(5) = 7$.

$q = 6$: Theorem 19 and Corollary 22 imply that $\mathcal{G}$ cannot be of type I or IV. So $\mathcal{G}$ is of type II or III. If it is of type II, then by Theorem 19, $|\mathcal{G}| = 2q - 4 = 8$. If it is of type III, then by Theorem 20, $|\mathcal{G}| \leq 2q - 4 = 8$. Therefore $F(6) = 8$ (see Figure 37).

$q = 7$: Corollary 22 imply that $\mathcal{G}$ cannot be of type IV. So $\mathcal{G}$ is of type I, II, or III. If it is of type I, then from the proof of Theorem 19, we must have $x_2 = x_3 = y_3 = y_1 = z_1 = z_2 = 2$ and $n_1 = n_2 = n_3 = 3$, and so $|\mathcal{G}| = 9$. If it is of type II, then by Theorem 19, $|\mathcal{G}| = 2q - 4 = 10$. If it is of type III, then by Theorem 20, $|\mathcal{G}| \leq 2q - 4 = 10$. Therefore $F(7) = 10$ (see Figure 38).

Observation. It is easy to verify that, for $8 \leq q \leq 14$, we have $3q+6-\lceil 6\sqrt{q+1}\, \rceil = 2q - 4$.

$8 \leq q \leq 14$: Corollary 22 implies that $\mathcal{G}$ cannot be of type IV. If it is of type I, then from Corollary 32 and the above observation, $|\mathcal{G}| \leq 2q-4$. If it is of type II, then by Theorem 19, $|\mathcal{G}| = 2q - 4$. If it is of type III, then by Theorem 20, $|\mathcal{G}| \leq 2q - 4$. Therefore $F(q) = 2q - 4$.

$q = 15$: $\mathcal{G}$ cannot be of type IV. If it is of type I, then $|\mathcal{G}| = |\mathcal{G}_0| = 27$ since $q = r^2 + 2r$ with $r = 3$. If it is of type II, then by Theorem 19, $|\mathcal{G}| = 2q - 4 = 26$. If it is of type III, then by Theorem 20, $|\mathcal{G}| \leq 2q - 4 = 26$. Therefore $F(15) = 27$.

$q = 16$: $\mathcal{G}$ cannot be of type IV. If it is of type I, then $|\mathcal{G}| \leq 3q + 6 - \lceil 6\sqrt{q+1}\, \rceil = 29$. If it is of type II, then by Theorem 19, $|\mathcal{G}| = 2q - 4 = 28$. If it is of type III, then by Theorem 20, $|\mathcal{G}| \leq 2q - 4 = 28$. Using an exhaustive search strategy shows that there is no graph of type I with 29 nodes. Therefore $F(16) = 28$. $\square$

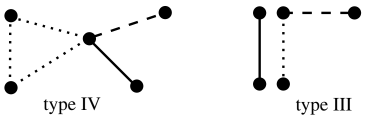


FIG. 35. *Maximal graphs when $q = 4$.*

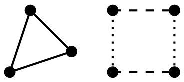


FIG. 36. *A maximal graph of type* III *when $q = 5$.*

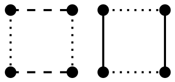


FIG. 37. *A maximal graph of type II when $q = 6$.*

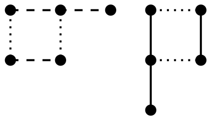


FIG. 38. *A maximal graph of type* II *when $q = 7$.*

7. Searching algorithm for maximal graphs. From Theorem 23, we know that, for $q \geq 17$, maximal graphs are of type I. Theorem 33 shows that the size of the maximal graphs is at most 2 different from the size of $\mathcal{G}_0$. In the rest of this section, we develop an efficient ($O(q^3)$) algorithm that finds the actual values of the defining parameters of the three 2-color parts of the maximal graphs.

LEMMA 35. *Suppose n_1, n_2, n_3, x_2, x_3, y_3, y_1, z_1, z_2 maximize the sum $n_1 + n_2 + n_3$ subject to the conditions of Definition* 26*. Then*

- $n_1 = y_1 z_1$ *or* $n_1 + x_2 + x_3 = q$,
- $n_2 = z_2 x_2$ *or* $n_2 + y_3 + y_1 = q$,
- $n_3 = x_3 y_3$ *or* $n_3 + z_1 + z_2 = q$,
- $\frac{1}{6}\sqrt{q+1} - 2 < x_2,\ x_3,\ y_3,\ y_1,\ z_1,\ z_2 < 6\sqrt{q+1}$

and, if $n_1 + x_2 + x_3 < q$, then

- $n_1 = y_1 z_1$,
- $n_2 + y_3 + y_1 = q$,
- $n_3 + z_1 + z_2 = q$.

Proof. If $n_1 < y_1 z_1$ and $n_1 + x_2 + x_3 < q$, then we can replace n_1 with $n_1 + 1$ to obtain a larger value for $n_1 + n_2 + n_3$ while the conditions are still satisfied. This is a contradiction.

Thus $n_1 = y_1 z_1$ or $n_1 + x_2 + x_3 = q$. Similarly, we have $n_2 = z_2 x_2$ or $n_2 + y_3 + y_1 = q$; also $n_3 = x_3 y_3$ or $n_3 + z_1 + z_2 = q$.

We have $n_1 + n_2 + n_3 = F(q) \geq 3q + 6 - 6\lceil \sqrt{q+1}\ \rceil > 3q - 6\sqrt{q+1}$. Moreover, $n_1 + x_2 + x_3 + n_2 + y_3 + y_1 + n_3 + z_1 + z_2 \leq 3q$, and so $x_2 + x_3 + y_3 + y_1 + z_1 + z_2 < 6\sqrt{q+1}$. Therefore, $x_2,\ x_3,\ y_3,\ y_1,\ z_1,\ z_2 < 6\sqrt{q+1}$.

We have $n_1 + n_2 + n_3 > 3q - 6\sqrt{q+1}$, and n_1, n_2, $n_3 \leq q$, which implies n_1, n_2, $n_3 > q - 6\sqrt{q+1}$. Since $y_1 z_1 \geq n_1 > q - 6\sqrt{q+1}$ and $z_1 < 6\sqrt{q+1}$ we have $y_1 > (q - 6\sqrt{q+1})/6\sqrt{q+1} = \frac{1}{6}\sqrt{q+1} - 1 - 1/(6\sqrt{q+1}) > \frac{1}{6}\sqrt{q+1} - 2$. Similarly,

Table 1
$F(q)$ and $|\mathcal{G}_0|$ for $1 \le q \le 48$.

q	$F(q)$	$\lvert\mathcal{G}_0\rvert$	q	$F(q)$	$\lvert\mathcal{G}_0\rvert$	q	$F(q)$	$\lvert\mathcal{G}_0\rvert$	q	$F(q)$	$\lvert\mathcal{G}_0\rvert$
1	1		13	22	21	25	49	49	37	79	79
2	2		14	24	24	26	52	52	38	82	81
3	4		15	27	27	27	54	54	39	84	84
4	5		16	28	28	28	57	57	40	87	87
5	7		17	31	31	29	60	60	41	90	90
6	8		18	33	33	30	62	62	42	92	92
7	10		19	36	36	31	64	63	43	94	93
8	12	12	20	38	38	32	67	66	44	97	96
9	14	13	21	40	39	33	69	69	45	99	99
10	16	16	22	42	42	34	72	72	46	102	102
11	18	18	23	45	45	35	75	75	47	105	105
12	20	20	24	48	48	36	76	76	48	108	108

we can prove $x_2,\ x_3,\ y_3,\ y_1,\ z_1,\ z_2 > \frac{1}{6}\sqrt{q+1} - 2$.

If $n_1 + x_2 + x_3 < q$, then $n_1 = y_1 z_1$. If we also have $n_2 + y_3 + y_1 < q$, then we can replace n_2 with $n_2 + 1$ and x_2 with $x_2 + 1$, while the conditions remain satisfied, to obtain a larger value for the sum $n_1 + n_2 + n_3$. This is a contradiction. So we must have $n_1 + x_2 + x_3 = q$; $n_3 + z_1 + z_2 = q$ is proved similarly. □

From Lemma 35, we can assume that $n_2 + y_3 + y_1 = q$ and $n_3 + z_1 + z_2 = q$, and we have either $n_1 = y_1 z_1$ or $n_1 + x_2 + x_3 = q$. Therefore to find n_1, n_2, n_3, x_2, x_3, y_3, y_1, z_1, z_2, we have the following $O(q^3)$ algorithm.

```
Initialize Max = $|\mathcal{G}_0|$.
Initialize $(n_1, n_2, n_3, x_2, x_3, y_3, y_1, z_1, z_2)$ = that of $\mathcal{G}_0$.
FOR  $x_2,\ x_3,\ y_3,\ y_1,\ z_1,\ z_2$ in the range $\frac{1}{6}\sqrt{q+1} - 2$ to $6\sqrt{q+1}$
LOOP
  $n_2 = q - (y_3 + y_1)$,
  $n_3 = q - (z_1 + z_2)$,
  $n_1 = q - (x_2 + x_3)$, or $y_1 z_1$.
  IF  conditions satisfied, then
    IF  $n_1 + n_2 + n_3 >$ Max,
      Max $= n_1 + n_2 + n_3$
      Save $(n_1, n_2, n_3, x_2, x_3, y_3, y_1, z_1, z_2)$
    END IF
  END IF
END LOOP
Output $F(q)$ = Max.
Output $(n_1, n_2, n_3, x_2, x_3, y_3, y_1, z_1, z_2)$.
```

Table 1 displays the outputs $F(q)$ of the algorithm in comparison with the values $|\mathcal{G}_0|$ for $1 \le q \le 48$.

Appendix. Proof of Theorem 27. We prove by contradiction. Suppose that, for some integer $q \ge 5$, there exist positive integers n_1, n_2, n_3, x_2, x_3, y_3, y_1, z_1, z_2, satisfying all conditions stated in Definition 26, and $n_1 + n_2 + n_3 > 3q + 6 - 6\sqrt{q+1}$. Since $n_1 + n_2 + n_3 + x_2 + x_3 + y_3 + y_1 + z_1 + z_2 \le 3q$, it follows that $x_2 + x_3 + y_3 + y_1 + z_1 + z_2 < 6\sqrt{q+1} - 6$. Hence, $x_2 + x_3 + y_3 + y_1 + z_1 + z_2 \le \lceil 6\sqrt{q+1}\ \rceil - 7$.

Therefore, at least one of the values $x_2+y_3+z_1$ and $x_3+y_1+z_2$ must be less than or equal to $(\lceil 6\sqrt{q+1}\,\rceil - 7)/2$. Without loss of generality, assume that

$$x_3+y_1+z_2 \le \frac{\lceil 6\sqrt{q+1}\,\rceil - 7}{2}. \tag{1}$$

Since $x_3+y_1+z_2 \ge 6$, q cannot be less than or equal to 8, so $q \ge 9$.

Consider $s = n_1+n_2+n_3+\lambda_1(y_1z_1-n_1)+\lambda_2(z_2x_2-n_2)+\lambda_3(x_3y_3-n_3)+(1-\lambda_1)(q-n_1-x_2-x_3)+(1-\lambda_2)(q-n_2-y_3-y_1)+(1-\lambda_3)(q-n_3-z_1-z_2)$, where λ_1, λ_2, and λ_3 are real numbers satisfying

$$\lambda_1 y_1 = 1-\lambda_3, \tag{2}$$

$$\lambda_2 z_2 = 1-\lambda_1, \tag{3}$$

$$\lambda_3 x_3 = 1-\lambda_2. \tag{4}$$

We have

$$\lambda_1 = \frac{1+z_2x_3-z_2}{1+y_1z_2x_3},\ \lambda_2 = \frac{1+x_3y_1-x_3}{1+y_1z_2x_3},\ \lambda_3 = \frac{1+y_1z_2-y_1}{1+y_1z_2x_3}.$$

Clearly, $0<\lambda_1,\lambda_2,\lambda_3<1$. Therefore, $s \ge n_1+n_2+n_3 > 3q+6-6\sqrt{q+1}$. We will derive a contradiction by showing that $s \le 3q+6-6\sqrt{q+1}$.

Claim 1. $\lambda_1+\lambda_2+\lambda_3 \le 1$.

Proof. From (2), (3), and (4), we have $3-(\lambda_1+\lambda_2+\lambda_3) = \lambda_1y_1+\lambda_2z_2+\lambda_3x_3 \ge 2(\lambda_1+\lambda_2+\lambda_3)$. Therefore, $\lambda_1+\lambda_2+\lambda_3 \le 1$.

Claim 2. $\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3} < 3q+3-6\sqrt{q+1}$.

Proof. Since for any $x,y>0$, if $\frac{1+x}{1+y}>1$, then $\frac{1+x}{1+y}<\frac{x}{y}$, we have

$$\frac{1}{\lambda_1} = \frac{1+y_1z_2x_3}{1+z_2x_3-z_2} < \frac{y_1z_2x_3}{z_2x_3-z_2} = y_1\frac{x_3}{x_3-1}.$$

Using similar inequalities for λ_2 and λ_3, we have

$$\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3} < y_1\frac{x_3}{x_3-1}+z_2\frac{y_1}{y_1-1}+x_3\frac{z_2}{z_2-1}.$$

If $y_1=z_2=x_3=2$, then $\lambda_1=\lambda_2=\lambda_3=\frac{1}{3}$ and $\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3} = 9 < 3q+3-6\sqrt{q+1}$.

If at least one of y_1, z_2, x_3 is greater than 2, say, $y_1>2$, then

$$\begin{aligned}\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3} &< y_1\frac{x_3}{x_3-1}+z_2\frac{y_1}{y_1-1}+x_3\frac{z_2}{z_2-1} \le 2y_1+\frac{3}{2}z_2+2x_3 \\ &\le 2(y_1+z_2+x_3)-1 \le \lceil 6\sqrt{q+1}\,\rceil - 8 \qquad \text{(by (1))}.\end{aligned}$$

It is easy to show that if $q\ge 9$, then $\lceil 6\sqrt{q+1}\,\rceil < 3q+11-6\sqrt{q+1}$. Therefore, $\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3} < 3q+3-6\sqrt{q+1}$.

Claim 3. $s \le 3q+6-6\sqrt{q+1}$.

Proof. We have

$$\begin{aligned} s = {}& \frac{1}{\lambda_1}[\lambda_1y_1-(1-\lambda_3)][\lambda_1z_1-(1-\lambda_2)] - \frac{1}{\lambda_1}(1-\lambda_2)(1-\lambda_3) \\ &+\frac{1}{\lambda_2}[\lambda_2z_2-(1-\lambda_1)][\lambda_2x_2-(1-\lambda_3)] - \frac{1}{\lambda_2}(1-\lambda_3)(1-\lambda_1) \\ &+\frac{1}{\lambda_3}[\lambda_3x_3-(1-\lambda_2)][\lambda_3y_3-(1-\lambda_1)] - \frac{1}{\lambda_3}(1-\lambda_1)(1-\lambda_2) \\ &+q(3-\lambda_1-\lambda_2-\lambda_3). \end{aligned}$$

From (2), (3), and (4), we have

$$\begin{aligned} s &= -\frac{1}{\lambda_1}(1-\lambda_2)(1-\lambda_3) - \frac{1}{\lambda_2}(1-\lambda_3)(1-\lambda_1) - \frac{1}{\lambda_3}(1-\lambda_1)(1-\lambda_2) \\ &\quad + q(3-\lambda_1-\lambda_2-\lambda_3) \\ &= -\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3}\right) + (\lambda_1+\lambda_2+\lambda_3)\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3}\right) - 3 \\ &\quad - \left(\frac{\lambda_2\lambda_3}{\lambda_1}+\frac{\lambda_3\lambda_1}{\lambda_2}+\frac{\lambda_1\lambda_2}{\lambda_3}\right) + q(3-\lambda_1-\lambda_2-\lambda_3). \end{aligned}$$

From Cauchy's inequality, we have

$$\frac{\lambda_2\lambda_3}{\lambda_1}+\frac{\lambda_3\lambda_1}{\lambda_2} \geq 2\lambda_3,\ \frac{\lambda_3\lambda_1}{\lambda_2}+\frac{\lambda_1\lambda_2}{\lambda_3} \geq 2\lambda_1,\ \frac{\lambda_1\lambda_2}{\lambda_3}+\frac{\lambda_2\lambda_3}{\lambda_1} \geq 2\lambda_2,$$

so

$$\frac{\lambda_2\lambda_3}{\lambda_1}+\frac{\lambda_3\lambda_1}{\lambda_2}+\frac{\lambda_1\lambda_2}{\lambda_3} \geq \lambda_1+\lambda_2+\lambda_3.$$

Therefore,

$$\begin{aligned} s &\leq (3q-3) - (q+1)(\lambda_1+\lambda_2+\lambda_3) - \left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3}\right) \\ &\quad +(\lambda_1+\lambda_2+\lambda_3)\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3}\right) \\ &= 2q-4+[1-(\lambda_1+\lambda_2+\lambda_3)]\left[q+1-\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3}\right)\right] \\ &= 2q-4+\frac{1}{q+1}[q+1-(q+1)(\lambda_1+\lambda_2+\lambda_3)]\left[q+1-\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3}\right)\right] \\ &\leq 2q-4+\frac{1}{4(q+1)}\left[2q+2-(q+1)(\lambda_1+\lambda_2+\lambda_3)-\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3}\right)\right]^2. \end{aligned}$$

From *Claims* 1 and 2, we have

$$\begin{aligned} &2q+2-(q+1)(\lambda_1+\lambda_2+\lambda_3)-\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3}\right) \\ &\geq q+1-\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3}\right) \\ &> q+1-(3q+3-6\sqrt{q+1}) = -(2q+2-6\sqrt{q+1}) \end{aligned}$$

and, from Cauchy's inequality,

$$\begin{aligned} &2q+2-(q+1)(\lambda_1+\lambda_2+\lambda_3)-\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3}\right) \\ &= 2q+2-(q+1)\lambda_1-\frac{1}{\lambda_1}-(q+1)\lambda_2-\frac{1}{\lambda_2}-(q+1)\lambda_3-\frac{1}{\lambda_3} \\ &\leq 2q+2-2\sqrt{q+1}-2\sqrt{q+1}-2\sqrt{q+1} = 2q+2-6\sqrt{q+1}. \end{aligned}$$

Therefore,

$$s \leq 2q-4+\frac{1}{4(q+1)}(2q+2-6\sqrt{q+1})^2 = 3q+6-6\sqrt{q+1}.$$

This completes the proof of Theorem 27. □

Acknowledgments. The authors would like to thank Greg Doherty for several very helpful comments and suggestions.

REFERENCES

[1] D. Boneh and J. Shaw, *Collusion-secure fingerprinting for digital data*, IEEE Trans. Inform. Theory, 44 (1998), pp. 1897–1905.

[2] B. Chor, A. Fiat, and M. Naor, *Tracing traitors*, in Advances in Cryptology (CRYPTO '94), Lecture Notes in Comput. Sci. 839, 1994, pp. 257–270.

[3] H. D. L. Hollmann, J. H. van Lint, J. Linnartz, and L. M. G. M. Tolhuizen, *On codes with the identifiable parent property*, J. Combin. Theory Ser. A, 82 (1998), pp. 121–133.

[4] S. Roman, *Coding and Information Theory*, Springer-Verlag, Berlin, New York, 1992.

[5] J. N. Staddon, D. R. Stinson, and R. Wei, *Combinatorial properties of frameproof and traceability codes*, IEEE Trans. Inform. Theory, 47 (2001), pp. 1042–1049.

[6] D. R. Stinson and R. Wei, *Combinatorial properties and constructions of traceability schemes and frameproof codes*, SIAM J. Discrete Math., 11 (1998), pp. 41–53.

SIAM J. Discrete Math.
Vol. 17, No. 4, pp. 571–581

RADIUS THREE TREES IN GRAPHS WITH LARGE CHROMATIC NUMBER*

H. A. KIERSTEAD[†] AND YINGXIAN ZHU[‡]

Abstract. A class Γ of graphs is χ-bounded if there exists a function f such that $\chi(G) \leq f(\omega(G))$ for all graphs $G \in \Gamma$, where χ denotes chromatic number and ω denotes clique number. Gyárfás and Sumner independently conjectured that, for any tree T, the class $\mathrm{Forb}(T)$, consisting of graphs that do not contain T as an induced subgraph, is χ-bounded. The first author and Penrice showed that this conjecture is true for any radius two tree. Here we use the work of several authors to show that the conjecture is true for radius three trees obtained from radius two trees by making exactly one subdivision in every edge adjacent to the root. These are the only trees with radius greater than two, other than subdivided stars, for which the conjecture is known to be true.

Key words. forbidden induced subgraph, template, radius three tree, chromatic number, clique number

AMS subject classification. 05C15

DOI. 10.1137/S0895480198339869

1. Introduction.

For an integer n, let $[n]$ denote the set $\{1, \ldots, n\}$. The complete graph on n vertices is denoted by K_n, and the complete bipartite graph with s vertices in one part and t vertices in the other part is denoted by $K_{s,t}$. For a graph $G = (V, E)$ and a subset $W \subseteq V$, let $G[W]$ denote the subgraph of G induced by W. For another graph H, we say that G *induces* H if H is isomorphic to an induced subgraph of G. When G is clear from the context, we may sometimes write W for $G[W]$. The clique size of G is denoted by $\omega(G)$ and the chromatic number of G is denoted by $\chi(G)$. As above, we may write $\omega(W)$ and $\chi(W)$ for $\omega(G[W])$ and $\chi(G[W])$ when G is clear from the context.

A class Γ of graphs is said to be *χ-bounded* if there exists a function f such that $\chi(G) \leq f(\omega(G))$ for every graph $G \in \Gamma$. For a graph H, let $\mathrm{Forb}(H)$ denote the class of graphs that do not contain an induced copy of H. In this paper, we study the following conjecture which is due independently to Gyárfás and Sumner.

Conjecture 1.1 (Gyárfás [1] and Sumner [8]). *For every tree T,* $\mathrm{Forb}(T)$ *is χ-bounded.*

The conjecture, if true, is essentially as strong as possible. Since there are graphs with arbitrarily large chromatic number and girth, $\mathrm{Forb}(H)$ is not χ-bounded if H contains a cycle. The following easy proposition reduces the problem for forests to the conjecture.

Proposition 1.2. *Let F be a forest with connected components $T_1, \ldots, T_m$. Then* $\mathrm{Forb}(F)$ *is χ-bounded iff* $\mathrm{Forb}(T_i)$ *is χ-bounded for all $i \in [m]$.*

The strongest partial result concerning Conjecture 1.1 is the following theorem of the first author and Penrice. It is proved by generalizing a "template" technique first introduced (but not named) by Gyárfás, Szemerédi, and Tuza [3].

*Received by the editors June 3, 1998; accepted for publication (in revised form) November 28, 2003; published electronically May 17, 2004.
http://www.siam.org/journals/sidma/17-4/33986.html

[†]Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287 (kierstead@math.la.asu.edu).

[‡]Department of Mathematics, Utah Valley State College, Orem, UT 84058 (zhuy@uvsc.edu).

THEOREM 1.3 (Kierstead and Penrice [5]). *For every tree T with radius at most two,* Forb (T) *is χ-bounded.*

The only other trees T for which Forb (T) was known to be χ-bounded were obtained from the following theorem of Scott. Let Forb* (H) be the class of graphs that do not contain an induced subdivision of H. Note that if T is a subdivision of a star, then Forb* $(T) =$ Forb (T).

THEOREM 1.4 (Scott [7]). *If T is a tree, then* Forb* (T) *is χ-bounded.*

COROLLARY 1.5 (Scott [7]). *If T is a subdivision of a star, then* Forb (T) *is χ-bounded.*

A natural goal is to try to extend Theorem 1.3 to radius three trees. The main result of this paper is the following theorem that takes a significant step in this direction. Our proof uses all the techniques of [5], techniques from [7] and [4], as well as several new ideas.

THEOREM 1.6. *If T is the tree obtained from a tree with radius two by making exactly one subdivision in every edge adjacent to the root, then* Forb (T) *is χ-bounded.*

We shall need the following theorem of Rödl.

THEOREM 1.7 (Rödl [6]). *For all trees T and positive integers t,* Forb $(T) \cap$ Forb $(K_{t,t})$ *is χ-bounded.*

This paper is organized as follows. In the remainder of this section we introduce our notation. In section 2 we prove a local coloring result based on the work of Scott [7]. In section 3 we present several general coloring techniques that will be used later. In section 4 we review the necessary results on templates from [5]. The proof of Theorem 1.6 is given in sections 5 and 6. In section 5 we describe a partitioning of the vertices of a graph based on 1- and 2-neighborhoods of templates, and in section 6 we use this partition to color the vertices of the graph. The novelty of this work involves adapting the template technique of Kierstead and Penrice [5] to take advantage of the new local colorings made available by Scott's ideas [7]. This adaptation is not straightforward. The partition presented in section 5 is considerably more delicate than the partition used in [5]. Moreover, the coloring techniques required to take advantage of this partition in section 6 are much harder than those used in [5].

For $r, k \geq 2$, let P_r denote the *path* on r vertices and S_k denote the *star* on $k+1$ vertices. The root of P_r is defined to be one of its leaves, while the root of S_k is defined to be its unique nonleaf. The (r,k)*-broom* $B_{r,k}$ is formed by identifying the nonroot leaf of P_r with the root of S_k. The other leaf of P_r is the root of $B_{r,k}$. A *spider* is a subdivision of a star or, alternatively, the result of identifying the roots of a collection of disjoint paths. A *spider with toes S* is the result of identifying the roots of a collection of disjoint brooms. The root of a spider with toes S is the new vertex obtained by this identification. Let $R_{d,k}$ be the spider with toes obtained by identifying the roots of d copies of $B_{3,k}$. Let R_k denote $R_{k,k}$. Our main result is equivalent to the statement that Forb (R_k) is χ-bounded for all k. The distance $d(v,w)$ between two vertices v and w is the number of edges in the shortest path from v to w.

Let $G = (V,E)$ be a graph and $W \subseteq V$. The *neighborhood* $N(W)$ of W is defined by $N(W) = \{v \in V - W : vw \in E \text{ for some } w \in W\}$. The *$i$-neighborhood* $S(W,i)$ of W is the set of vertices x such that the shortest path from x to a vertex in W is i. For $v \in V$, we may write $N(v)$ instead of $N(\{v\})$ and $S(v,i)$ instead of $S(\{v\},i)$. The *radius* of G is the least integer r such that $V \subseteq \bigcup_{i=0}^{r} S(v,i)$ for some vertex v. We say that two sets of vertices U and W are adjacent if some vertex in U is adjacent to some vertex in W. Also a vertex v is said to be adjacent to W if $\{v\}$ is adjacent to W. Let

$\mathrm{Ram}\,(a,b)$ be a Ramsey function such that any graph on $\mathrm{Ram}\,(a,b)$ vertices contains a clique with a vertices or an independent set with b vertices. A function $f : V \to S$ is a proper S-coloring of G if $f\,(x) \neq f\,(y)$ whenever $xy \in E$. If $f : A \to B$ is a function and $S \subseteq A$, then the image $\{f\,(s) : s \in S\}$ of S under f is denoted by $f\,[S]$.

2. 2-neighborhoods. When arguing by induction on the clique size, we can assume that there is a bound on the chromatic number of the neighborhood of any vertex. A key step in our current proof is showing that we can also assume that the 2-neighborhood of any vertex has bounded chromatic number. This is the content of Lemma 2.2 below. The proof uses Lemma 2.1, which is a special case of a more general lemma of the first author [4]. Its short proof is given for completeness. We also need a 2-neighborhood partition (defined in the first paragraph of the proof of Lemma 2.2) and other ideas due to Scott [7].

LEMMA 2.1 (Kierstead [4]). *Let $H = (V,E)$ be a graph with $\chi\,(H) = c > hk$. If $S \subseteq V$ satisfies both $\chi\,(S) \leq h$ and $\chi\,(V - S) < \chi\,(H)$, then there exists a vertex $v \in S$ such that v is adjacent to k vertices in $H - S$.*

Proof. Suppose to the contrary that every vertex $v \in S$ is adjacent to less than k vertices in $H - S$. By hypothesis, there exist a proper $[c-1]$-coloring f of $V - S$ and a proper $[h]$-coloring g of S. We shall obtain a contradiction by showing that f can be extended to a proper $[c-1]$-coloring of H as follows. For $v \in S$, let $f\,(v) = ih + g\,(v)$, where i is the least natural number such that v is not adjacent to any vertex $w \in V - S$ with $f\,(w) = ih + g\,(v)$. Since v has less than k neighbors in $V - S$, we have $i < k$. So $f\,(v) \in [hk] \subseteq [c-1]$. To check that f is proper, consider $u, v \in S$ and $w \in V - S$ with v adjacent to u and w. By the construction, $f\,(v) \neq f\,(w)$ and $f\,(v) \neq f\,(u)$ since

$$f\,(v) \mod h = g\,(v) \neq g\,(u) = f\,(u) \mod h. \qquad \square$$

LEMMA 2.2. *There exists a natural number $m = m\,(h,k)$ such that, for every graph $H = (V,E) \in \mathrm{Forb}\,(R_k)$, if $\chi\,(N\,(u)) < h$ for every $u \in V$, then $\chi\,(S\,(v,2)) < m$ for every $v \in V$.*

Proof. Let F_i be the forest consisting of i components, each of which is isomorphic to the star S_k. By Proposition 1.2 and either Theorem 1.3 or Corollary 1.5, there exists a function f such that if G is a graph with $\chi\,(G) > f\,(\omega\,(G))$, then G contains an induced copy of the forest F_{hk}. Define a function g recursively by

$$g\,(0) = 0 \text{ and } g\,(d+1) = g\,(d) + f\,(h) + h^2k + hd\,(2k+3)\,.$$

Next, let $m(h,k) = \max\{f(h), g(hk)\}$. We shall finish the proof by showing that if $\chi\,(S\,(v,2)) > m\,(h,k)$, then H contains an induced copy of R_k.

Suppose that H satisfies the hypothesis of the lemma. Let v be any vertex in V and set $Y_0 = N\,(v)$, $X_0 = S\,(v,2)$, and $U_0 = \emptyset$. First, we recursively define a partition $\{Z_1, \ldots, Z_a\}$ of X_0, subsets $Y_1 \supseteq \cdots \supseteq Y_a$, and functions $\psi, \zeta_1, \psi_1, \ldots, \zeta_i, \psi_a$. Consider a positive integer i and suppose that Y_j and Z_j have already been defined for $j \in [i-1]$. Set $U_i = \emptyset$ if $i = 1$; otherwise, set $U_i = \bigcup_{j<i} Z_j$. Set $X_i = X_0 - U_i$. If $X_i = \emptyset$, then set $a = i - 1$ and stop; otherwise, let Y_i be a minimal subset of Y_{i-1} such that $N\,(x) \cap Y_i \neq \emptyset$ for all $x \in X_i$. By the minimality of Y_i, for each $y \in Y_i$ there exists $z \in X_i$ such that $N\,(z) \cap Y_i = \{y\}$. Let $\zeta_i\,(y)$ be such a z. Then $\zeta_i : Y_i \to X_i$ is an injection. Let Z_i be the range of ζ_i. So $\zeta_i : Y_i \to Z_i$ is a bijection. Let ψ_i be the inverse of ζ_i. So the set M_i of edges between Y_i and Z_i is the perfect matching $\{z\psi_i\,(z) : z \in Z_i\}$. This completes the recursive definition. Finally, define $\psi : X_0 \to Y_0$ by $\psi\,(z) = \psi_i\,(z)$ if $z \in Z_i$.

Case 1. $\chi(Z_i) > f(h)$ for some $i \in [a]$. Since $\omega(Z_i) \leq \max_{u \in Z_i} \chi(N(u)) < h$, there exists a subset $W \subseteq Z_i$ that induces F_{hk}. Let A be the hk-set of roots of components of F_{hk} and $B = \psi_i[A]$. Since $B \subseteq N(v)$ and $\chi(N(v)) < h$, there exists an independent k-subset $B' \subseteq B$. Let $A' = \zeta_i[B']$ and W' be the union of the components of W that have roots in A'. Since M_i is a perfect matching, it follows that $\{v\} \cup B' \cup W'$ induces R_k with root v.

Case 2. $\chi(Z_i) \leq f(h)$ for all $i \in [a]$. We first show by induction on d that if $\chi(U_i) > g(d)$, then there exist $W_d \subseteq U_i$ and $Q_d \subseteq Y_1$ such that W_d induces F_d, and $W_d \cup Q_d \cup \{v\}$ induces a subgraph G_d such that $G_d - E(Q_d)$ is isomorphic to $R_{d,k}$ with root v. The base step $d = 0$ is trivial, so consider the induction step $d = c + 1$. Suppose $\chi(U_i) > g(d)$. Let j be the least integer such that $\chi(U_j) > g(c)$. Since $U_j = U_{j-1} \cup Z_{j-1}$ and $\chi(Z_{j-1}) \leq f(h)$, we have $\chi(U_j) \leq g(c) + f(h)$. Thus

$$\chi(U_i - U_j) > h^2k + hc(2k+3).$$

Let W_c and Q_c be the sets whose existence is guaranteed by the induction hypothesis. Note that $|W_c| = c(k+1)$ and $|Q_c| = c$. Let $A = N(W_c \cup Q_c) \cap (U_i - U_j)$, $B = N(W_c) \cap Y_{j+1}$, $C = N(B) \cap (U_i - U_j)$, and $S = U_i - (U_j \cup A \cup C)$. Then $S \cup \psi[S]$ is not adjacent to $W_c \cup Q_c$. So it suffices to show that there exist $x \in S$ and a k-set $I \subseteq S \cap N(x)$ such that $\{x, \psi(x)\} \cup I$ induces $B_{2,k}$.

Suppose that $z \in W_c$. Then $z \in Z_m$ for some $m \in [j-1]$. Thus $\psi(z)$ is the only vertex in Y_m adjacent to z. Since $Y_{j+1} \subseteq Y_m$, at most one vertex in Y_{j+1} is adjacent to z. It follows that $|B| \leq |W_c| = c(k+1)$. Since $A \cup C \subseteq N(W_c \cup Q_c \cup B)$ and neighborhoods of vertices are h-colorable, it follows that $\chi(A \cup C) \leq h|W_c \cup Q_c \cup B| \leq hc(2k+3)$. Thus

$$\chi(S) \geq \chi(U_i - U_j) - \chi(A \cup C) > h^2k + hc(2k+3) - hc(2k+3) = h^2k.$$

Let $S' \subseteq S$ be minimal subject to the condition that $\chi(S') > h^2k$. Let $z \in S'$ and $y = \psi(z)$. By Lemma 2.1, there exists $x \in N(y)$ such that $|N(x) \cap (S' - N(y))| \geq hk$. Since $N(v)$ is h-colorable, there exists an independent k-subset $I \subseteq N(x) \cap (S' - N(y))$. Let $W_d = W_c \cup \{x\} \cup I$ and $Q_d = Q_c \cup \{y\}$. Then $G_d - E(Q_d)$ is isomorphic to $R_{d,k}$ with root v.

Now suppose that $\chi(X_0) > g(hk)$. Let W_{hk}, Q_{hk}, and G_{hk} be as above. Since $Q_{hk} \subseteq N(v)$ is h-colorable, there exists an independent k-subset $J \subseteq Q_{hk}$. Thus G_{hk} contains an induced copy of R_k with root v. □

3. Coloring techniques. A graph $G = (V, E)$ is *d-degenerate* if there exists an ordering $v_1 < v_2 < \cdots < v_n$ of V such that $|\{j \in [i-1] : v_jv_i \in E\}| \leq d$ for all $i \in [n]$. The *maximum average degree* of G is $\text{MAD}(G) = \max_{H \subseteq G} \frac{2|E(H)|}{|V(H)|}$. For a digraph D, let $\Delta^+(D)$ be the maximum outdegree of D. The next three lemmas are well known.

LEMMA 3.1. *Every graph* $G = (V, E)$ *is* $\text{MAD}(G)$*-degenerate.*

Proof. We argue by induction on $|V|$. The base step $|V| = 1$ is trivial, so consider the induction step. Let v be a vertex of G with minimum degree. Then $d(v) \leq \text{MAD}(G)$ and $\text{MAD}(G - v) \leq \text{MAD}(G)$. By the induction hypothesis, $G - v$ is $\text{MAD}(G)$-degenerate. Putting v at the end of an ordering of $V - v$ witnessing this produces an ordering witnessing that G is $\text{MAD}(G)$-degenerate. □

LEMMA 3.2. *If* G *is a* d*-degenerate graph, then* $\chi(G) \leq d + 1$.

Proof. Use first-fit to color the vertices of G in an order witnessing that G is d-degenerate. □

LEMMA 3.3. *Any directed graph* G *satisfies* $\chi(G) \leq 2\Delta^+(G) + 1$.

Proof. Since $|E(H)| \leq |V(H)| \Delta^{+}(G)$ for any subgraph $H \subseteq G$, it follows that $\mathrm{MAD}(G) \leq 2\Delta^{+}(G)$. By Lemma 3.1, G is $2\Delta^{+}(G)$-degenerate, and so by Lemma 3.2, we are done. □

Lemma 3.4. *Let $G = (V, E)$ be a graph and g be a positive integer. Suppose that $\{(A_i, B_i) : i \in I\}$ is a family of ordered pairs of subsets of V such that $|\{j \in I - \{i\} : A_i$ is adjacent to $B_j\}| < g$ for all $i \in I$. Then there exists a subset $J \subseteq I$ such that $|J| \geq \frac{|I|}{2g}$ and A_i is not adjacent to B_j for all distinct $i, j \in J$.*

Proof. Define a digraph $D = (I, A)$ by $i \to j$ iff $i \neq j$ and A_i is adjacent to B_j. We are looking for a large independent subset of I. By Lemma 3.3, $\chi(D) < 2g$ since $\Delta^{+}(D) < g$. Thus I contains an independent subset of size $\frac{|I|}{2g}$. □

In applications, A_i and B_i are not necessarily disjoint and may be identical.

Lemma 3.5. *Let $G = (V, E)$ be a graph and $\{I_j : j \in [p]\}$ be a partition of V into independent sets. Suppose that*

(1) *$|\{j \in [i-1] :$ some vertex in I_i is adjacent to at least k vertices in $I_j\}| < a_1$ for all $i \in [p]$, and*

(2) *$|\{j : x$ is adjacent to $I_j\}| < a_2$ for all $x \in V$.*

Then $\chi(G) < 2a_1a_2k$.

Proof. Consider a digraph $D = ([p], A)$ defined by $i \to j$ iff some vertex in I_i is adjacent to at least k vertices in I_j. By (1), $\Delta^{+}(D) < a_1$, and so by Lemma 3.3 $\chi(D) < 2a_1$. Let f be a proper $[2a_1]$-coloring of D. Let $g : V \to [a_2k]$ such that, for all $v \in I_i$ and $w \in I_j$, if v is adjacent to w and v is not adjacent to k vertices in I_j, then $g(v) \neq g(w)$. This is possible by (2). Define $h : V \to [2a_1] \times [a_2k]$ by $h(v) = (f(i), g(v))$, where $v \in I_i$. Clearly h is a proper coloring of G. □

4. Templates. We shall need the concept of a *k-template* introduced by the first author and Penrice in [5]. Let k be a positive integer and $G = (V, E)$ be a graph with $\omega(G) = n$. A k-template in G is a pair $\Gamma = (X, Y)$ of subsets of V such that

(1) there exists an integer i with $2 \leq i \leq n$ such that $G[Y]$ is a complete i-partite graph with $k + (n - i + 1)(k - 1)$ vertices in each part, and

(2) $X \subseteq \{v \in V - Y : v$ has less than k nonneighbors in each part of $Y\}$.

The *index* $I(\Gamma)$ of the k-template $\Gamma = (X, Y)$ is defined by $I(\Gamma) = (i, |X|)$, where i is the number of parts of $G[Y]$. We order the indices lexicographically by $(i, |X|) < (i', |X'|)$ iff $i < i'$ or $i = i'$ and $|X| < |X'|$. Let the index $I(G)$ of a graph G be the maximum index among all the k-templates of G; if this collection is empty, then $I(G) = (0, 0)$. Consider a template (X, Y) and a vertex $v \in Y$. We say that v is *rarely adjacent* to Y if v is adjacent to Y, but v is adjacent to less than k vertices in each part of Y. We say that v is *usually adjacent* to Y if v is nonadjacent to less than k vertices in each part of Y. We say that v is *often adjacent* to Y if v is adjacent to Y, but it is neither rarely adjacent nor usually adjacent to Y. Let $O(Y) = \{v \in V - Y : v$ is often adjacent to $Y\}$. Note that the vertices of X are usually adjacent to Y. The following lemmas from [5] show the existence of templates (Lemma 4.1), bound the size of maximum templates (Lemma 4.2), and show why maximum templates are useful (Lemma 4.3). Except for Lemma 4.1, which is an immediate corollary of Theorem 1.7, they have easy direct proofs.

Lemma 4.1 (Kierstead and Penrice [5]). *For all trees T and integers n, k, if G has a sufficiently large chromatic number, then G induces either T, or K_{n+1}, or a k-template.*

Lemma 4.2 (Kierstead and Penrice [5]). *Let G be a graph with $\omega(G) = n$. Let $\Gamma = (X, Y)$ be a k-template of G, where Y is i-partite. If $|X| > r(k, i)$, where*

$r(k,i) = \mathrm{Ram}\left(n+1, (k+(n-i)(k-1))\,2^{i(k+(n-i+1)(k-1))}\right)$, *then there exists a k-template $\Gamma' = (X', Y')$ such that $Y' \subseteq Y \cup X$ and Y' is a complete $(i+1)$-partite graph. Thus if $I(G) = (i, |X|)$, then $i \leq n$ and $|X| < r(k,i)$.*

LEMMA 4.3 (Kierstead and Penrice [5]). *Suppose $\Gamma = (X, Y)$ is a k-template of G, w is adjacent, but not rarely adjacent, to Y, and v is not usually adjacent to Y. If v is adjacent to $\{w\} \cup Y$, then there exists $D \subseteq \{w\} \cup Y$ such that $D \cup \{v\}$ induces $B_{2,k}$ with root v. Moreover, if v is adjacent to Y, then we can choose $D \subseteq Y$.*

Finally, we will need one new lemma on templates.

LEMMA 4.4. *Suppose $\Gamma = (X, Y)$ is a k-template of G. Let v be a vertex of $G - Y$ such that v is rarely adjacent to Y. Then there exists $D \subseteq Y$ such that $D \cup \{v\}$ induces $B_{3,k}$ with root v.*

Proof of Lemma 4.4. Since each part of Y has at least $2k-1$ vertices and v is rarely adjacent to Y, there exist vertices $u_0, u_1, \ldots, u_k$ in some part of Y such that v is adjacent to u_i iff $i = 0$. Also, there exists w in some other part of Y that is not adjacent to v. Clearly $v, u_0, w, u_1, \ldots, u_k$ induce $B_{3,k}$. □

5. The partition. We now begin the proof of our Theorem 1.6. Note that if T' is an induced subgraph of T, then $\mathrm{Forb}(T') \subseteq \mathrm{Forb}(T)$. Thus it suffices to show that for arbitrarily large k, $\mathrm{Forb}(R_k)$ is χ-bounded. Fix k. We shall argue by induction on the pair (n, I), ordered lexicographically, that there exists a function $b(n, I)$ such that, for any graph $G \in \mathrm{Forb}(R_k)$, if $\omega(G) \leq n$ and $I(G) \leq I$, then $\chi(G) \leq b(n, I)$. If $\omega(G) < 2$, the result is trivial, and if $I(G) < (2, 0)$, it follows from Lemma 4.1. So for the rest of the paper, let $G = (V, E)$ be any graph in $\mathrm{Forb}(R_k)$ with $\omega(G) = n \geq 2$ and $I(G) = I \geq (2, 0)$. Recall that, by Lemma 4.2, for any maximum k-template $\Gamma = (X, Y)$, both $|X|$ and $|Y|$ are bounded by a function of n (since k is fixed). We begin by recursively defining a family of disjoint subsets $\{X_i, Y_i, A_i, B_i : i \in [p]\} \cup \{L\}$.

Let $V_1 = V$ and suppose that we have defined X_j, Y_j, A_j, B_j, and V_{j+1} for all $j \in [i-1]$. If $(\omega(V_i), I(V_i)) < (n, I)$, then set $p = i - 1$ and $L = V_i$. Otherwise, let $\Gamma = (X_i, Y_i)$ be a k-template with index I in V_i, $A_i = V_i \cap O(Y_i)$, $B_i = (V_i - X_i) \cap N(Y_i \cup O(Y_i))$, and $V_{i+1} = V_i - (X_i \cup Y_i \cup A_i \cup B_i)$. This completes the definition of L and the set $\{X_i, Y_i, A_i, B_i\colon i \in [p]\}$. Let $X = \bigcup_{i\in[p]} X_i$, $Y = \bigcup_{i\in[p]} Y_i$, $A = \bigcup_{i\in[p]} A_i$, and $B = \bigcup_{i\in[p]} B_i$.

We shall need the following lemmas about this partition.

LEMMA 5.1. *$I(L) < I$.*

Proof. The proof is immediate from the halting condition. □

LEMMA 5.2. *For all $i < j \leq p$, the sets $Y_i \cup O(Y_i)$ and $X_j \cup Y_j \cup A_j \cup B_j$ are not adjacent.*

Proof. Since $X_j \cup Y_j \cup A_j \cup B_j \subseteq V_{i+1} = V_i - (Y_i \cup N(Y_i) \cup N(O(Y_i)))$, it follows that $Y_i \cup O(Y_i)$ and $X_j \cup Y_j \cup A_j \cup B_j$ are nonadjacent. □

Let $c = k + 2$.

LEMMA 5.3. *Let $v \in V$ and $J = \{j \in [p] : v$ is adjacent to $Y_j\}$. Then $|J| < c$. Moreover, if $J_3 = \{j \in [p] : v$ is usually adjacent to $Y_j\}$, then $|J_3| \leq 1$.*

Proof. Partition J into $J = J_1 \cup J_2 \cup J_3$, where

$$J_1 = \{j \in [p] : v \text{ is rarely adjacent to } Y_j\} \text{ and}$$
$$J_2 = \{j \in [p] : v \text{ is often adjacent to } Y_j\}.$$

Suppose $|J_1| \geq k$. By Lemma 4.4, for each $j \in J_1$, there exists $D_j \subseteq Y_j$ such that $\{v\} \cup D_j$ induces $B_{3,k}$. Thus, using Lemma 5.2, $\{v\} \cup \bigcup_{i\in J_1} D_j$ induces R_k, which is a contradiction. So $|J_1| < k$. Suppose $J_2 \neq \emptyset$. Let i be the least index in J_2. Then

$v \in O(Y_i)$. By Lemma 5.2, v is not adjacent to Y_j for $j > i$. So $|J_2| \leq 1$. Suppose that $i, j \in J_3$. Then by Lemma 5.2, $v \notin Y$. Then both $(X_i \cup \{v\}, Y_i)$ and $(X_j \cup \{v\}, Y_j)$ are k-templates in G. Since both (X_i, Y_i) and (X_j, Y_j) are maximum in G, it follows that $v \in X_i \cap X_j$. Thus $i = j$ and $J_3 \leq 1$. So $|J| < c = k + 2$. □

Let $d = c + \mathrm{Ram}(n, 2ck)$.

Lemma 5.4. *Let $v \in V$ and $J = \{j : v$ is adjacent to $Y_j \cup A_j\}$. Then $|J| < d$.*

Proof. Suppose that $|J| \geq d$. We shall obtain a contradiction by showing that G induces R_k. Let $J_0 = \{j \in J : v$ is not adjacent to $Y_j\}$. By Lemma 5.3, $|J_0| \geq \mathrm{Ram}(n, 2ck)$. For each $j \in J_0$, let $a_j \in A_j$ be adjacent to v. Since $\omega(G) = n$, there exists $J_1 \subseteq J_0$ such that

$$|J_1| = 2ck \text{ and } \{a_i : i \in J_1 \text{ is independent}\}.$$

By Lemma 5.3, $|\{i \in [p] : a_j$ is adjacent to $Y_i\}| < c$ for all $j \in J_1$. Thus by Lemma 3.4 applied to the family $\{(\{a_j\}, Y_j) : j \in J_1\}$, there exists a k-subset $I \subseteq J_1$ such that for all distinct $j, i \in I$, a_j is not adjacent to Y_i. By Lemma 4.3, for every $i \in I$ there exists $D_i \subseteq Y_i$ such that $D_i \cup \{a_i\}$ induces $B_{2,k}$ with root a_i. By Lemma 5.2, $\{v\} \cup \{a_i : i \in I\} \cup \bigcup_{i \in I} D_i$ induces R_k, which is a contradiction. □

Let $e_1 = d + \mathrm{Ram}(n, 4dk)$.

Lemma 5.5. *Let $v \in V$ and $J = \{j \in [p] : v$ is adjacent to $Y_j \cup A_j \cup B_j\}$. Then $|J| < e_1$.*

Proof. Suppose that $|J| \geq e_1$. We shall obtain a contradiction by showing that G induces R_k. Let $J_0 = \{j \in J : v$ is adjacent to neither Y_j nor $A_j\}$. By Lemma 5.4, $|J_0| \geq \mathrm{Ram}(n, 4dk)$. For each $j \in J_0$, let $b_j \in B_j$ be adjacent to v. Since $\omega(G) = n$, there exists $J_1 \subseteq J_0$ such that $|J_1| = 4dk$ and $\{b_j : j \in J_1\}$ is independent. For all $j \in J_1$, there exists $z_j \in Y_j \cup O(Y_j)$ such that z_j is adjacent to b_j. If possible, choose $z_j \in Y_j$. Otherwise, by the construction, $z_j \in A_j$. Regardless of whether $z_j \in A_j$ or $z_j \in Y_j$, there exists an independent k-set $D_j \subseteq Y_j$ such that $D_j \cup \{z_j, b_j\}$ induces $B_{2,k}$ with root b_j: if $z_j \in Y_j$, this follows from Lemma 4.3; otherwise, b_j is not adjacent to Y_j and it follows from the fact that $z_j \in A_j$ and so z_j is often adjacent to Y_j. By Lemmas 5.3 and 5.4, for all $i \in J_1$ we have

$$|\{j \in J_1 : \{b_i, z_i\} \text{ is adjacent to } z_j \cup D_j\}| < 2d.$$

Thus by Lemma 3.4 applied to the family $\{(\{b_j, z_{j,}\}, \{z_j \cup D_j\}) : j \in J_1\}$ there exists a k-subset $I \subseteq J_1$ such that $\{b_i, z_i\}$ is not adjacent to $\{z_j\} \cup D_j$ for all distinct $i, j \in I$. Then $\{v\} \cup \bigcup_{i \in I} (\{b_i, z_i\} \cup D_i)$ induces R_k. □

6. The coloring.

We shall complete the proof by showing that the chromatic number of each of A, B, X, Y, and L is bounded by a function of n.

Lemma 6.1. *$\chi(L)$ is bounded by a function of n.*

Proof. Trivially, $\omega(L) \leq \omega(G)$, and by Lemma 5.1, $I(L) < I$. Thus $(\omega(L), I(L)) < (n, I)$, and so by the induction hypothesis, $\chi(L)$ is bounded by a function of n. □

Lemma 6.2. *$\chi(Y)$ is bounded by a function of n.*

Proof. The proof follows immediately from the definition of templates that $\chi(Y_i) \leq n$ for all $i \in [p]$. By Lemma 5.2, Y_i is not adjacent to Y_j for distinct $i, j \in [p]$. Thus $\chi(Y) \leq n$. □

Lemma 6.3. *$\chi(X)$ is bounded by a function of n.*

Proof. Since each template (X_i, Y_i) has the same index, each X_i has the same cardinality, say t. By Lemma 4.2, t is bounded by a function of n. Partition X into $\{W_j : j \in [t]\}$ so that $|W_j \cap X_i| = 1$ for all $j \in [t]$ and $i \in [p]$. Let $f = c + 4ck$ and

$Q = \{u\} \cup \{v_j, w_j : j \in [f]\}$ be a spider with root u, where u is adjacent to each v_j and each v_j is adjacent to w_j. Using Theorem 1.3, it suffices to show that $W_m \in \operatorname{Forb}(Q)$, and thus $\chi(W_m)$ is bounded by a function of n for all $m \in [t]$.

Fix $m \in [t]$ and set $W = W_m$. Suppose W induces Q. Without loss of generality, let $Q \subseteq W$. We shall obtain a contradiction by showing that a subset of $Q \cup Y$ induces R_k with root u. For $j \in [f]$, define $g(j)$ by $v_j \in X_{g(j)}$ and $h(j)$ by $w_j \in X_{h(j)}$. Let $J = \left\{j \in [f] : u \text{ is not adjacent to } Y_{h(j)}\right\}$. Then $|J| \geq 4ck$ by Lemma 5.3. We claim that, for each $j \in J$, there exists $D_j \subseteq Y_{h(j)} \cup \{w_j\}$ such that $D_j \cup \{v_j\}$ induces $B_{2,k}$ with root v_j: since v_j is usually adjacent to $Y_{g(j)} \neq Y_{h(j)}$, by Lemma 5.3, v_j is not usually adjacent to $Y_{h(j)}$. Also w_j is usually adjacent to $Y_{h(j)}$ since $w_j \in X_{h(j)}$. Thus by Lemma 4.3, there exists $D_j \subseteq Y_{h(j)} \cup \{w_j\}$ such that $D_j \cup \{v_j\}$ induces $B_{2,k}$. By Lemma 5.3 and the fact that $\{v_j, w_j\}$ is not adjacent to w_i, we have $|\{i \in J : \{v_j, w_j\} \text{ is adjacent to } D_i\}| < 2c$ for all $j \in J$. Thus by Lemma 3.4 applied to the family $\{(\{v_j, w_j\}, D_j) : j \in J\}$, there exists a k-subset $I \subseteq J$ such that for all distinct $i, j \in I$, it is not the case that $\{v_j, w_j\}$ is adjacent to D_i. Thus by Lemma 5.2, $\{u\} \cup \bigcup_{i \in I} (D_i \cup \{v_i\})$ induces R_k. □

LEMMA 6.4. *$\chi(A \cup B)$ is bounded by a function of n.*

Proof. First we show that $\chi(A_i \cup B_i)$ is bounded by a function of n for all $i \in [p]$. Let $Y_i = \{y_{i,j} : j \in [|Y_i|]\}$. The vertices of $A_i \cup B_i$ can be partitioned into $2|Y_i|$ sets $B_{i,j}$ and $B^*_{i,j}$ such that

$$B_{i,j} \subseteq N(y_{i,j}) \text{ and } B^*_{i,j} \subseteq N(O(Y_i) \cap N(y_{i,j})) \cap S(y_{i,j}, 2).$$

Clearly $\omega(N(y_{i,j})) < \omega(G)$. So by the induction hypothesis, $\chi(B_{i,j})$ is bounded by a function of n. By Lemma 2.2 and the induction hypothesis, $\chi\left(B^*_{i,j}\right)$ is bounded by a function of n. Since $|Y_i|$ is bounded by a function of n, our claim is proved. Moreover, we can insist that for each color class S there exists $y \in Y_i$ such that $S \subseteq S(y,1) \cup S(y,2)$. Thus it suffices to prove that if $C = \bigcup_{i \in [p]} C_i$, where C_i is an independent subset of $(A_i \cup B_i) \cap (S(y_i,1) \cup S(y_i,2))$ for some $y_i \in Y_i$, then $\chi(C)$ is bounded by a function of n.

We say that a family $\mathcal{F}$ of subsets of C is *distinguishing* if all $i \in [p]$ satisfy $|\{S \in \mathcal{F} : C_i \cap S \neq \emptyset\}| \leq 1$. Notice that, if $\mathcal{F}$ is a distinguishing family, then by Lemma 5.5 every vertex is adjacent to at most e_1 elements of $\mathcal{F}$.

Define a new digraph $H = ([p], F)$ as follows. Let $e_2 = f_0(f_2 + 1)$, $f_0 = e_1 + f_1$, $f_1 = 2e_1(k+2)k$, $f_2 = f_0 \operatorname{Ram}(f_3, n)$, $f_3 = 2f_4$, and $f_4 = 2e_1(k+2)k$. In particular, $e_2 \geq 2e_1 + 2f_1 \geq f_4 + (k+1)e_1$. We say that (u, M) is a *witness* for (i,j) iff $u \in C_i$, $M \subseteq [p] - \{i,j\}$, $|M| = e_2$, and for every $m \in M$, there exists $w \in C_j \cap N(u)$ such that w is adjacent to at least k vertices in C_m. Let F be the set of ordered pairs (i,j) for which there exists a witness.

PROPOSITION 6.5. *The outdegree of H is less than e_2.*

Proof. Suppose that the outdegree of $i \in [p]$ is at least e_2. We shall obtain a contradiction by showing that G induces R_k. Let $J = \{j \in [p] : (i,j) \in F\}$. For each $j \in J$, let (u_j, M_j) be a witness for (i,j).

Case 1. There exist $v \in V$ and $J_0 \subseteq J$ such that $|J_0| = f_0$ and $\{u_j : j \in J_0\} \subseteq N(v)$. Let $K = \{j \in [p] : v \text{ is adjacent to } C_j\}$. Then $|K| \leq e_1$. Choose a subset $J_1 \subseteq J_0 - K$ with $|J_1| = f_1 = f_0 - e_1$. Also choose an injection $m : J_1 \to [p]$ so that $m(j) \in M_j - (K \cup J_1)$ and u_j is not adjacent to $C_{m(j)}$. This is possible since $|M_j| - (K \cup J_1) \geq e_2 - 2e_1 - f_1 \geq f_1$. Finally, for each $j \in J_1$ choose $w_j \in C_j \cap N(u_j)$ and a k-set $Z_j \subseteq C_{m(j)} \cap N(w_j)$. Then $\{v, u_j, w_j\} \cup Z_j$ induces $B_{3,k}$ with root v and $\{\{w_j\}, Z_j : j \in J_1\}$ is a distinguishing family. It follows that $\{\{w_j\} \cup Z_j : j \in J_1\}$ is also a distinguishing family. So by Lemma 3.4 applied to $\{(\{u_j, w_j\} \cup Z_j, \{w_j\} \cup Z_j) : j \in J_1\}$ there exists a subset $J_2 \subseteq J_1$ such that $|J_2| = \frac{f_1}{2(k+2)e_1} = k$ and $\{u_j, w_j\} \cup Z_j$ is not adjacent to $\{w_{j'}\} \cup Z_{j'}$ for all distinct $j, j' \in J_2$. Since C_i is independent and $\{u_j : j \in J_2\} \subseteq C_i$, it follows that $\{v\} \cup \bigcup_{j \in J_2} (\{u_j, w_j\} \cup Z_j)$ induces R_k.

Case 2. Case 1 fails. Let $J_2 = \{j \in J : y_i$ is not adjacent to $C_j\}$ and $U_2 = \{u_j : j \in J_2\}$. Using Lemma 5.5 and the case, $|U_2| \geq \frac{e_2 - e_1}{f_0} \geq f_2$, note that $U_2 \subseteq S(y_i, 2)$. Let $S \subseteq N(y_i)$ be a minimal set such that $U_2 \subseteq N(S)$ by the case $|S| \geq \frac{f_2}{f_0} = \operatorname{Ram}(f_3, n)$. Since $S \subseteq N(y_i)$ we have $\omega(S) < n$. Thus S contains an independent subset S_0 with $|S_0| = f_3$. Using the minimality of S, there exists an injection $j : S_0 \to J_2$ such that s is the unique element of S_0 that is adjacent to $u_{j(s)}$. For each $s \in S_0$ and $m \in M_{j(s)}$, there exists $w_{s,m} \in N(u_{j(s)}) \cap C_{j(s)}$ such that $w_{s,m}$ is adjacent to C_m. Let $W_s = \{w_{s,m} : m \in M_{j(s)}\}$. Since $W_s \subseteq C_{j(s)}$, it is independent. By Lemma 5.5, $|W_s| \geq \frac{e_2}{e_1} \geq k$. Let $S_1 = \{s \in S_0 : |W_s - N(s)| \geq k\}$.

Case 2*a*. $|S_1| \geq f_4$. For each $s \in S_1$, let Z_s be a k-subset of $W_s - N(s)$. Then $\{y_i, s, u_{j(s)}\} \cup Z_s$ induces $B_{3,k}$ with root y_i, and $\{y_i\} \cup \bigcup_{s \in S_1} \{s, u_{j(s)}\}$ induces a spider. Moreover, $\{Z_s : s \in S_1\}$ is a distinguishing family. Thus by Lemma 3.4 applied to the set $\{(\{s, u_{j(s)}\} \cup Z_s, Z_s)\}$ there exists $S_2 \subseteq S_1$ with $|S_2| = \frac{f_4}{2(k+2)e_1} = k$ such that $\{s, u_{j(s)}\} \cup Z_s$ and $Z_{s'}$ are not adjacent for all distinct $s, s' \in S_2$. Thus $\{y_i\} \cup \bigcup_{s \in S_2} (\{s, u_{j(s)}\} \cup Z_s)$ induces R_k.

Case 2*b*. $|S_1| < f_4$. Let $S_3 \subseteq S_0 - S_1$ with $|S_3| = f_4 = f_3 - f_4$. Then every $s \in S_3$ is adjacent to all but at most $k - 1$ vertices of W_s. Let $J_3 = \{j(s) : s \in S_3\}$. Let $M'_{j(s)} = \{m \in M_{j(s)} : w_{s,m} \in N(s)\}$. Then $\left|M'_{j(s)}\right| \geq e_2 - (k-1)e_1$. Choose an injection $m : S_3 \to [p]$ so that $m(s) \in M'_{j(s)} - J_3$ and neither s nor y_i is adjacent to $C_{m(s)}$. This is possible because

$$|S_3| = f_4 \leq e_2 - (k-1)e_1 - 2e_1.$$

Let Z_s be a k-subset of $N(w_{s,m(s)}) \cap C_{m(s)}$ and $T_s = \{w_{s,m(s)}\} \cup Z_s$. Then $\{s\} \cup T_s$ induces $B_{2,k}$ and $\{T_s : s \in S_3\}$ is a distinguishing family. When Lemma 3.4 is applied to $\{(\{s\} \cup T_s, T_s) : s \in S\}$, there exists a subset $S_4 \subseteq S_3$ such that $|S_4| = \frac{f_4}{2(k+2)e_1} = k$ and $\{s\} \cup T_s$ is not adjacent to $T_{s'}$ for distinct $s, s' \in S_4$. It follows that $\{y_i\} \cup S_4 \cup \bigcup_{s \in S_4} T_s$ induces R_k. □

By Lemma 3.3 and Proposition 6.5, we can properly color H with $2e_2$ colors. Thus it suffices to show that $C' = \bigcup_{i \in I} C_i$ has bounded chromatic number for any H-independent subset $I \subseteq [p]$. Partition C' into $\{N_i : i \in [e_1]\} \cup \{M_{e_1}\}$ recursively as follows. Let $M_0 = C'$. Now suppose that we have defined M_t. Call a vertex $v \in M_t$ *good* if there exist $i, j \in I$ with $i < j$ such that $v \in C_i$ and v is adjacent to at least k vertices in $C_j \cap M_t$. Let M_{t+1} be the set of all good vertices in M_t. Also let $N_{m+1} = M_m - M_{m+1}$, $Q = M_{e_1}$, and $Q_i = Q \cap C_i$. Using Lemma 5.5, $\operatorname{MAD}(N_m) \leq 2(k-1)e_1$; by Lemma 3.1, N_m is $(2(k-1)e_1)$-degenerate; and by Lemma 3.2, $\chi(N_m) \leq 2ke_1$ for all $m < e_1$. Thus it suffices to show that Q has bounded chromatic number. Define an auxiliary digraph $H' = (I, F')$ by $(i, j) \in F'$ iff $i < j$ and some vertex in Q_i is adjacent to at least $2k$ vertices in Q_j. By Lemmas 3.5 and 5.5, it suffices to prove Proposition 6.6 below. The following definitions are needed to state this proposition. They are presented in a way that emphasizes clarity in the proof of the proposition at the expense of clarity in the definition. Let $e_3 = ge_1$, $g = g_1(g' + 1)$, $g' = \operatorname{Ram}(g_0, n)$, $g_0 = 2e_1(k+1)k$, $g_1 = 2(k+1)e_2g_2$, $g_2 = g_3g_4$, $g_3 = 2ke_1g_4$, $g_4 = g_5g_7$, $g_5 = 2e_1(k+1)k$, $g_7 = g_8 + g_{10}$, $g_8 = 2k^2e_2$, $g_{10} = 2kg_{11}$, and $g_{11} = 2e_1k(k+1)$.

PROPOSITION 6.6. *The outdegree of H' is less than e_3.*

Proof. Suppose the outdegree of $i \in I$ is at least e_3 in H'. We shall obtain a contradiction by showing that G induces R_k. Let $J = \{j \in I : (i,j) \in F'\}$. Let U be a minimum subset of Q_i such that for all $j \in J$ there exists $u \in U$ such that u is adjacent to a $2k$-subset of Q_j. Then U is independent and, for all $u \in U$, there exists $h(u) \in J$ such that u is the unique vertex in U that is adjacent to at least $2k$ vertices in $Q_{h(u)}$. Let D_u be a $2k$-subset of $N(u) \cap Q_{h(u)}$. Then $\{D_u : u \in U\}$ is a distinguishing family. By Lemma 5.5, $|U| \geq g = \frac{e_3}{e_1}$. Let $O_i = (N(y_i) \cap O(Y_i)) \cup \{y_i\}$. Then $U \subseteq N(O_i)$. By Lemma 5.2, O_i is not adjacent to C_j for any $j > i$.

Case 1. No $s \in O_i$ is adjacent to more than g_1 vertices of U. Let $S \subseteq O_i$ be a minimum set such that $U \subseteq N(\{y_i\} \cup S)$. Then $|S| \geq g' = \frac{g}{g_1} - 1$, and for every vertex $s \in S$ there exists a neighbor $u_s \in N(s) \cap U$ such that s is the unique vertex in $\{y_i\} \cup S$ adjacent to u_s. So $\{y_i, s, u_s\} \cup D_{u_s}$ induces $B_{2,k}$. Also $\omega(S) < n$ since $S \subseteq N(y_i)$. Since $|S| \geq g' = \mathrm{Ram}(g_0, n)$, there exists an independent subset $S_0 \subseteq S$ such that $|S_0| = g_0$. Then $\{y_i\} \cup \bigcup_{s \in S_0} \{s, u_s\}$ induces a spider. Let $U_0 = \{u_s : s \in S_0\}$. By Lemma 3.4 applied to $\{(\{u_s\} \cup D_{u_s}, D_{u_s}) : s \in S_0\}$, there exists $S_1 \subseteq S_0$ with $|S_1| = \frac{g_0}{2e_1(k+1)} = k$ such that $\{u_s\} \cup D_{u_s}$ is not adjacent to $D_{u_{s'}}$ for all distinct $s, s' \in S_1$. It follows that $\{y_i\} \cup \bigcup_{s \in S_1} (\{s, u_s\} \cup D_{u_s})$ induces R_k.

Case 2. Some vertex $s \in O_i$ is adjacent to more than g_1 vertices of U. Let $U_1 = U \cap N(s)$. So $|U_1| \geq g_1$. We claim that for each vertex $u \in U_1$ there exist indices $l(u)$ and $m(u)$ with $i < l(u) < m(u)$ and a vertex $v_u \in C_{l(u)} \cap N(u)$ such that $\left|C_{m(u)} \cap N(v_u)\right| \geq k$ and $C_{m(u)} \cap N(u) = \emptyset$. To see this fix u and let $l(0)$ be the greatest index such that $M_{e_1} \cap C_{l(0)} \cap N(u)$ is nonempty. Now suppose we have defined an increasing sequence of indices $l(0), \ldots, l(t)$ such that $M_{e_1 - r} \cap C_{l(r)} \cap N(u)$ is nonempty for all $r \in [t]$. Let $v \in M_{e_1 - t} \cap C_{l(t)} \cap N(u)$. Let $l(t+1)$ be the largest index such that $l(t) < l(t+1)$ and $\left|M_{e_1 - t - 1} \cap C_{l(t+1)} \cap N(v)\right| \geq k$. The existence of $l(t+1)$ follows from the definition of $M_{e_1 - t}$. If $C_{l(t+1)} \cap N(v) = \emptyset$, then set $l(u) = l(t)$, $m(u) = l(t+1)$, and $v_u = v$. Otherwise continue. Eventually the process must terminate since, by Lemma 5.5, u is adjacent to less than e_1 sets C_j.

For each vertex $u \in U_1$, let Z_u be a k-subset of $C_{m(u)} \cap N(v_u)$. So $\{u, v_u\} \cup Z_u$ induces $B_{2,k}$. Since $s \in O_i$ and $i < l(u) < m(u)$, we have that s is not adjacent to $\{v_u\} \cup Z_u$. So $\{s, u, v_u\} \cup Z_u$ induces $B_{3,k}$. Since I is independent in H, every vertex in C' is adjacent to less than e_2 vertices in U_1: if $v \in C_h$ is adjacent to e_2 vertices in U, then (v, J_1) is a witness for (h, i), where $J_1 = \{h(u) : u \in U_1\}$. Thus by Lemma 3.4 applied to the family $\{(\{v_u\} \cup Z_u, \{u\}) : u \in U_1\}$ there exists a subset $U_2 \subseteq U_1$ with $|U_2| \geq g_2 = \frac{g_1}{2(k+1)e_2}$ such that u is not adjacent to $\{v_{u'}\} \cup Z_{u'}$ for distinct $u, u' \in U_2$.

Consider the set $M = \{m(u) : u \in U_2\}$. If $|M| \geq g_3$, then let U_3 be a minimal subset of U_2 such that for every $m \in M$ there exists $u \in U_3$ such that $m(u) = m$. Using the choice of U_2 we have that $|U_3| \geq g_3$ and $\{Z_u : u \in U_3\}$ is a distinguishing family. Thus by Lemma 3.4 applied to the family $\{(Z_u, Z_u) : u \in U_3\}$ there exists $U_4 \subseteq U_3$ with $|U_4| = g_4 = \frac{g_3}{2ke_1}$ such that Z_u is not adjacent to $Z_{u'}$ for distinct $u, u' \in U_4$. Otherwise there exists $m \in M$ such that $U_4 = \{u \in U_2 : m(u) = m\}$ has cardinality at least $g_4 = \frac{g_2}{g_3}$. Again Z_u is not adjacent to $Z_{u'}$ for distinct $u, u' \in U_4$, since $Z_u, Z_{u'} \subseteq C_m$, which is independent.

Consider the set $M' = \{l(u) : u \in U_4\}$. If $|M'| \geq g_5$, then let U_5 be a minimal subset of U_4 such that for every $l \in M'$ there exists $u \in U_5$ such that $l(u) = l$. Then $|U_5| \geq g_5$. Also $\{\{v_u\} : u \in U_5\}$ is a distinguishing family. Thus by Lemma 3.4 applied to the family $\{(\{v_u\} \cup Z_u, \{v_u\}) : u \in U_5\}$ there exists $U_6 \subseteq U_5$ with $|U_6| \geq \frac{g_5}{2(k+1)e_1} = k$ such that $Z_u \cup \{v_u\}$ is not adjacent to $v_{u'}$ for distinct $u, u' \in U_6$. Then

$\{s\} \cup \bigcup_{u \in U_6} (\{u, v_u\} \cup Z_u)$ induces R_k. Otherwise there exists $l \in M'$ such that $U_7 = \{u \in U_4 : l(u) = l\}$ has cardinality at least $g_7 = \frac{g_4}{g_5}$. Again $\{v_u : u \in U_6\}$ is independent. But we are not done. We would like to show that U_7 contains a k-set U' such that Z_u is not adjacent to $v_{u'}$ for all distinct $u, u' \in U'$.

Consider the set $U_8 = \{u \in U_7 : |N(v_u) \cap D_u| \geq k\}$. First suppose that $|U_8| \geq g_8$. Since I is independent in H, each vertex in C' is adjacent to less than e_2 vertices in $\{v_u : u \in U_7\}$. Thus, by Lemma 3.4 applied to the family $\{(Z_u, \{v_u\}) : u \in U_8\}$ there exists $U_9 \subseteq U_8$ with $|U_9| = \frac{g_8}{2ke_2} = k$ such that Z_u is not adjacent to $v_{u'}$ for all distinct $u, u' \in U_9$. Thus $\{s\} \cup \bigcup_{u \in U_9} (\{u, v_u\} \cup Z_u)$ induces R_k. Otherwise $U_{10} = U_7 - U_8$ has cardinality at least $g_{10} = g_7 - g_8$. Then for each $u \in U_{10}$ there exists a k-set $D'_u \subseteq D_u$ such that v_u is not adjacent to D'_u.

Suppose there exist a vertex $r \in \bigcup_{u \in U_{10}} Z_u$ and a subset $U_{11} \subseteq U_{10}$ such that $\{v_u : u \in U_{11}\} \subseteq N(r)$ and $|U_{11}| = g_{11}$. Since $\{D'_u : u \in U_{11}\}$ is a distinguishing family, by Lemma 3.4 applied to the family $\{(\{r\} \cup D'_u, D'_u) : u \in U_{11}\}$ there exists a subset $U_{12} \subseteq U_{11}$ with $|U_{12}| = \frac{g_{11}}{2(k+1)e_1} = k$ such that $\{r\} \cup D'_u$ is not adjacent to $D'_{u'}$ for all distinct $u, u' \in U_{11}$. It follows that $\{r\} \cup \bigcup_{u \in U_{12}} (\{v_u, u\} \cup D'_u)$ induces R_k. Otherwise, by Lemma 3.4 applied to the family $\{(Z_u, \{v_u\}) : u \in U_{10}\}$ there exists a subset $U_{13} \subseteq U_{10}$ with $|U_{13}| = \frac{g_{10}}{2g_{11}} = k$ and Z_u not adjacent to $v_{u'}$ for all distinct $u, u' \in U_{12}$. It follows that $\{s\} \cup \bigcup_{u \in U_{13}} (\{u, v_u\} \cup Z_u)$ induces R_k. □ □

This completes the proof of Theorem 1.6.

REFERENCES

[1] A. Gyárfás, *On Ramsey covering-numbers*, in Infinite and Finite Sets, Colloq. Math. Soc. János Bolyai 10, North-Holland, Amsterdam, 1975, pp. 801–816.

[2] A. Gyárfás, *Problems from the world surrounding perfect graphs*, Zastos. Mat., 19 (1987), pp. 413–441.

[3] A. Gyárfás, E. Szemerédi, and Z. Tuza, *Induced subtrees in graphs of large chromatic number*, Discrete Math., 30 (1980), pp. 235–244.

[4] H. A. Kierstead, *Classes of graphs that are not vertex Ramsey*, SIAM J. Discrete Math., 10 (1997), pp. 373–380.

[5] H. Kierstead and S. Penrice, *Radius two trees specify χ-bounded classes*, J. Graph Theory, 18 (1994), pp. 119–129.

[6] H. Kierstead and V. Rödl, *Applications of hypergraph coloring to coloring graphs not inducing certain trees*, Discrete Math., 150 (1996), pp. 187–193.

[7] A. D. Scott, *Induced trees in graphs of large chromatic number*, J. Graph Theory, 24 (1997), pp. 247–311.

[8] D. P. Sumner, *Subtrees of a graph and chromatic number*, in The Theory and Applications of Graphs (Kalamazoo, MI, 1980), G. Chartrand, Y. Alavi, D. L. Goldsmith, L. Lesniak-Foster, and D. L. Lick, eds., Wiley, New York, 1981, pp. 557–576.

SIAM J. Discrete Math.
Vol. 17, No. 4, pp. 582–595

MINIMIZING WIRELENGTH IN ZERO AND BOUNDED SKEW CLOCK TREES[*]

MOSES CHARIKAR[†], JON KLEINBERG[‡], RAVI KUMAR[§], SRIDHAR RAJAGOPALAN[§], AMIT SAHAI[†], AND ANDREW TOMKINS[§]

Abstract. An important problem in VLSI design is distributing a clock signal to synchronous elements in a VLSI circuit so that the signal arrives at all elements simultaneously. The signal is distributed by means of a clock routing tree rooted at a global clock source. The difference in length between the longest and shortest root-leaf path is called the *skew* of the tree. The problem is to construct a clock tree with zero skew (to achieve synchronicity) and minimal sum of edge lengths (so that circuit area and clock tree capacitance are minimized).

We give the first constant-factor approximation algorithms for this problem and its variants that arise in the VLSI context. For the zero skew problem in general metric spaces, we give an approximation algorithm with a performance guarantee of $2e$. For the L_1 version on the plane, we give an $(8/\ln 2)$-approximation algorithm.

Key words. clock routing, skew, wirelength minimization

AMS subject classifications. 68W20, 68W25, 68W35, 68W40

DOI. 10.1137/S0895480199352622

1. Introduction.

A fundamental problem in VLSI design is *clock routing*, i.e., distributing a clock signal to synchronous elements in a VLSI circuit so that the signal arrives at all elements simultaneously. The signal is distributed by means of a clock routing tree rooted at a global clock source. The difference in length between the longest and shortest root-leaf path is called the *skew* of the tree. To achieve synchronicity, the skew should be zero. This is a significant issue in VLSI design, as nonzero clock skew has been estimated to account for over 10% of overall system cycle time in some high-performance systems [3]. Though it is easy to produce zero skew clock routing trees (see, e.g., [4]), naive algorithms may produce trees that are expensive in terms of total *wirelength* (i.e., sum of the edge lengths in the tree), thereby increasing circuit area and clock tree capacitance. Thus, the ideal clock tree routing algorithm would produce a zero skew clock tree with minimal total wirelength.

This problem, well studied in the VLSI community [14, 13, 7, 6, 23, 17, 25, 5, 15, 19, 8], is precisely the following variant of the classical *Steiner tree problem*:

> Find a Steiner tree, with a distinguished root, so that the lengths of all the root-leaf paths are the same and the sum of the lengths of edges in the tree is minimized.

[*] Received by the editors February 17, 1999; accepted for publication (in revised form) September 2, 2003; published electronically May 17, 2004. A preliminary version of this paper appeared in *Proceedings of the* 10*th ACM–SIAM Symposium on Discrete Algorithms*, 1999, pp. 177–184.

http://www.siam.org/journals/sidma/17-4/35262.html

[†] Department of Computer Science, Princeton University, Princeton, NJ 08544 (moses@cs.princeton.edu, sahai@princeton.edu). This work was done while the authors were visiting IBM Almaden Research Center. The second author was also supported in part by a DOD/NDSEG fellowship and DARPA grant DABT63-96-C-0018.

[‡] Department of Computer Science, Cornell University, Ithaca, NY 14853 (kleinber@cs.cornell.edu). This work was done while the author was visiting IBM Almaden Research Center. The author was also supported in part by an Alfred P. Sloan Research Fellowship and by NSF Faculty Early Career Development Award CCR-9701399.

[§] IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120 (ravi@almaden.ibm.com, sridhar@almaden.ibm.com, tomkins@almaden.ibm.com).

While there are many proposed heuristics for attacking this problem and its variants (see, for instance, the papers cited above), there are no algorithms with nontrivial worst-case performance guarantees known. In this paper we give the first (constant-factor) approximation algorithms for constructing clock trees with zero skew (or a skew of at most a fixed bound), and wirelength as small as possible.

1.1. Clock routing problems. We focus on the following three versions of the (zero or bounded skew) clock routing problem.

1. **L_1 clock routing.** A clock signal must be distributed using horizontal and vertical wires on the plane from a source to a set of terminal points. The most common model of delay along a wire is the *linear model*, in which delay corresponds to length. Therefore the distance between points is exactly the L_1 distance. This is the standard formulation of the problem.
2. **Planar L_1 clock routing.** In general, the embedding of a clock tree may have intersecting wires since the terminals are usually placed first, and then two layers of metal are available for the horizontal and vertical wires of the clock tree. This crossing of wires, however, may necessitate the introduction of many *vias*, or connections between layers, which causes both additional unmodeled delay and attenuation of the clock signal. Therefore one requires a planar-embeddable clock tree [14]. We therefore consider a second version of the routing problem (under the L_1 metric on the plane) with the requirement that the resulting clock tree be a planar embedding.
3. **General metric space clock routing.** The above two versions model the clock routing problem for standard-cell or gate-array design methodologies, which have many small functional modules. In contrast to this, building-block design methodologies use larger functional blocks. These blocks are treated as obstacles and routing must be done in the spaces between blocks. The routing problem is formulated with respect to a graph, called the *channel intersection graph* (CIG) that represents the available routing area. In this model we can think of the terminals V as embedded in a metric space induced by the topology of the CIG. Therefore the third variant of the problem we study is routing a clock tree in an arbitrary metric space.

1.2. Preliminaries. We are given a metric space M with distance function d, and a set V of points in M that are designated as *terminals*. As is standard, we define a *Steiner tree* for V to be a tree in M that contains each terminal in V as a vertex. (The vertices of the tree other than the terminals are referred to as *Steiner vertices.*) We say that a *clock tree T* for V is a Steiner tree with a distinguished vertex r called the *root*, such that every terminal $v \in V$ occurs as a leaf of T. The tree has an associated length function d_T that assigns a length to every edge in T, subject to the restriction that $d_T(u, v) \geq d(u, v)$ (i.e., the tree is allowed to stretch distances).

Existing algorithms for clock routing in the L_1 plane make use of *snaking*, or wiggling an edge in order to lengthen it. Our definition of clock tree incorporates snaking by allowing $d_T(u, v) > d(u, v)$. Without this extension, no zero skew tree may exist. In our model, feasibility is no longer a concern—any tree whose leaves are exactly the terminals can be "snaked" to a higher-cost zero skew tree.

If the metric space is the L_1 plane, for instance, the length of an edge (u, v) in T is at least the L_1 distance between u and v. The *cost* of the tree T is the sum of the lengths of all the edges of T. For $v \in V$, let the length of the path from the root to v be $\ell_v = d_T(r, v)$. The *skew* of T is $\max_{u,v \in V} |\ell_u - \ell_v|$. If T has skew $= 0$, we call it a *zero skew (clock) tree* (ZST) and if T has skew at most s, we call it an *s-skew tree.* (Note that, if necessary, this definition can be modified to allow terminals to be internal vertices of the tree; in the plane, we can instead slightly displace the internal vertex from the terminal. For general possibly discrete metric spaces, we allow multiple points in the tree to correspond to the same point of the metric space.)

Formally, the *zero* (resp., *bounded*) *skew clock tree problem* is stated as follows.

> Given a set V of terminals in a metric space M, find the minimum cost zero skew tree (resp., tree with skew at most s for a given bound s) for V.

When M is the L_1 plane, we refer to the L_1 variants of these problems. As discussed earlier, intersecting wires in the embedding might cause additional unmodeled delays. This motivates the *planar* variants of the above problem, where the tree T must be planar-embeddable (i.e., have no crossing edges).

The bounded skew clock tree problem is easily seen to be **NP**-complete by setting the skew to infinity so that the problem becomes the classical Steiner tree problem. The same reduction implies that the problem has no approximation scheme in general metrics unless **P** = **NP**. The zero skew problem is also **NP**-complete for general metric spaces. To our best knowledge, the hardness question of the planar zero skew problem is yet unsolved.

We will also refer to these problems as the zero or bounded skew clock routing problems.

1.3. Our results. For the ZST problem in general metric spaces, we give an approximation algorithm with a performance guarantee of $2e \approx 5.44$. We then give an approximation algorithm for the bounded skew clock routing problem in general metric spaces with a performance guarantee of 16.1065. Finally we give an $(8/\ln 2)$-approximation algorithm for the planar ZST problem and a constant-factor approximation algorithm for the planar-embeddable bounded skew clock routing problem.[1]

1.4. Organization. Section 2 discusses some related work in clock routing. Section 3 presents a general lower bound for the optimal solution to the problem. This is used to obtain approximation guarantees for our algorithms. Section 4 (resp., section 5) gives the approximation algorithms for the zero (resp., bounded) skew clock routing problems. Section 6 presents an approximation algorithm for the planar ZST problem, and section 7 presents an approximation algorithm for the planar-embeddable bounded skew clock routing problems. Section 8 discusses the hardness of the ZST problem.

2. Related work. Algorithms for clock tree constructions come in two flavors—those that guarantee zero skew and others that attempt to minimize the skew. Notice, however, that the aim is typically to minimize total wirelength.

The book by Kahng and Robins [14] contains a detailed account of many of the algorithms for clock tree constructions and several experimental results. The main emphasis on many of the algorithms, however, is to obtain practical solutions (which perform well on standard benchmarks, which in turn may or may not represent the average-case problem instance) rather than obtain solutions which have worst-case performance guarantees. We review the most relevant algorithms below.

2.1. Minimizing clock skew. In [19], given a floor plan of modules, a scheme to identify an entry point is presented. The optimal layout of the clock lines from

[1] Since the appearance of this paper, the approximation factors for the zero and bounded skew clock routing problems in general metric spaces have been improved to 4 and 14, respectively [24].

the source to the entry points is determined by an exhaustive search (of course, with some pruning). No theoretical guarantees on the performance of the algorithm are given.

In [13], the authors obtain a clock routing scheme consisting of Manhattan segments with constraints (like blockages) on the routing layers. They obtain a divide-and-conquer algorithm which produces total wirelength of $1.5\sqrt{n}$ for n terminals distributed randomly on a uniform grid. Contrasting this with the largest possible wirelength of $\sqrt{n}+1$ for a rectilinear Steiner tree for the same distribution, they conclude that, on average, their algorithm is a 3/2-approximation algorithm when compared to the minimum rectilinear Steiner tree.

Another algorithm for minimizing skew and wirelength based on matching is given in [5, 15]. They construct a binary tree using geometric matching and show that for cell-based designs, the total wirelength of their routing tree is on average, within a constant factor of the wirelength in an optimal Steiner tree. Their experiments suggest that the skew is near-zero on average.

2.2. Zero clock skew. An exact zero skew clock routing algorithm using the Elmore delay model is presented in [21, 22]. The zero skew is obtained by a bottom-up hierarchical approach via a zero skew merging of the recursive solutions. The main emphasis is on experimental results.

A two-step approach to obtaining zero skew while simultaneously minimizing wirelength is pioneered in [4]. In this, the authors present the *Deferred Merge Embedding* (DME) algorithm, which embeds any given connection topology to create a zero skew clock tree. The wirelength is optimal for linear delay. The connection topology is generated by a top-down balanced bipartition (BB) approach. Though the DME algorithm can be shown to produce the optimal tree for a given topology, the BB approach is essentially a heuristic and has no performance guarantees.

3. A lower bound.

We first demonstrate a lower bound on the cost of the s-skew tree in any metric space. Let T be any rooted s-skew tree on the set V of terminals, with root-leaf path length (i.e., the radius) R'. Since T has skew at most s, the length of every root-leaf path is at least $R' - s$.

We define the *level* of a vertex $p \in T$ to be its distance in the tree from the root (so the root is at level 0). Consider some level $x \in [0, R' - s]$. If there are m vertices at level x in T, then the m spheres of radius $R' - x$ centered at these vertices must cover all the terminals of V. This observation can be converted into a lower bound as follows. Let $n_V(R)$ be the minimum number of spheres of radius R needed to cover the terminals V. When the set is apparent from context, we suppress the subscript V.

Let Δ be the diameter of the set of terminals V, and let R^* be the minimum value of x such that $n(x) = 1$ (thus $R^* > \Delta/2$). Note that $R' \geq R^*$. Then the cost of the minimum cost s-skew tree must be at least

$$\int_0^{R'-s} n(R' - x)dx = \int_s^{R'} n(R)dR \geq \int_s^{R^*} n(R)dR.$$

This lower bound, and the special case for $s = 0$, will be essential to analyzing our algorithms.

4. An approximation algorithm for general metric spaces.

In this section, we present a $2e$-approximation algorithm for the ZST problem in general arbitrary metric spaces (assuming that snaking is valid). The algorithm is randomized but can

```
Algorithm Connect-Centers:
Initialize:  $R := R_0; u_0 := s; U_0 := V; \bar{U} := \{U_0\}; \bar{G} := \emptyset; T := \emptyset; i := 0; R_{\text{old}} := \Delta/2$.
repeat until $i = |V|$
  $S := V; i = 0$.
  repeat until $S = \emptyset$
    pick $g_i$ arbitrarily from $S$.
    let $G_i$ be all vertices in $S$ within distance $2R$ from $g_i$.
    let $S := S \setminus G_i; i := i + 1$.
  for $j = 0$ to $i - 1$ do
    let $k$ be such that $g_j \in U_k$.
    add an edge from $g_j$ to $u_k$ of cost exactly $2R_{\text{old}}$ to $T$.
  $R_{\text{old}} := R; R := R/r; \bar{U} := \bar{G}; \bar{G} := \emptyset$.
Output $T$.
```

FIG. 1. *Algorithm* Connect-Centers.

be derandomized easily. We place Steiner vertices on top of terminals from V. For ease of language, when we talk of using a terminal as an internal point in the tree, we mean to place a Steiner vertex at that terminal and use the Steiner vertex as the internal vertex in the tree.

Our algorithm repeatedly partitions the set of vertices to construct the tree. The partitioning proceeds by greedily placing balls of a certain radius $2R$ and grouping all vertices in the same ball together. To obtain more and more refined partitions, the process repeats with balls of smaller radii. We denote by r the factor by which the radii of balls decrease in each successive refinement of the partitioning process. We will describe our algorithm for any value of r and choose a specific (optimal) value for r at the end.

Algorithm Connect-Centers. Let Δ be the diameter of V. The algorithm first picks an arbitrary vertex s to be the root of the tree, and then chooses an initial partitioning radius $2R_0$ as follows. Let t be chosen uniformly at random from $[0, 1]$, and set $R_0 = (\Delta/2)\cdot\exp(-t \ln r)$. The algorithm then proceeds as in Figure 1. At each point in the construction, we take an existing partition of the vertices $\bar{U}$ and refine it to $\bar{G}$. ($\bar{G}$ is not necessarily a strict refinement of $\bar{U}$.) Each set $G_i \in \bar{G}$ has a distinguished member g_i with the property that every $v \in G_i$ has $d(v, g_i) \leq 2R$. Similarly each $U_i \in \bar{U}$ has a member u_i such that every $v \in U_i$ has $d(v, u_i) \leq 2R_{\text{old}} = 2rR$. The tree we construct is denoted by T.

Remark. The algorithm as presented in Figure 1 is only weakly polynomial. But, by constraining R to be $\leq R_0/n^2$, we can obtain a strongly polynomial algorithm at the expense of $O(1/n)$ additive factor in the performance ratio.

Analysis. It is immediate from the description of the algorithm that it will return a ZST, since each vertex is reached after the same number of levels, and the edges in each level are of identical cost. To analyze the cost of the tree produced by this algorithm, we observe the following lemma.

LEMMA 4.1. *Each time a new partition G is created the number of sets returned in the partition is at most $n_V(R)$.*

Proof. Let $G = \{G_0, \ldots, G_{m-1}\}$. We induct on m. If $m = 1$, there is nothing to prove. Otherwise, consider the $n = n_V(R)$ sets $S_1, S_2, \ldots, S_n$ of radius R that cover all the terminals V. Let S_j be the set that contains g_0. Since G_0 contains all vertices within radius $2R$ from g_0, it must contain all of S_j. Let $V' = V \setminus G_0$. Now, certainly, $n_V(R) \geq 1 + n_{V'}(R)$. But by induction, since the sets $G_1, G_2, \ldots, G_{m-1}$ are the result of a valid execution of the partitioning algorithm on V', it follows that $m - 1 \leq n_{V'}(R)$, and so the claim follows. Note that the claim also follows from the standard analysis for the p-center problem [10, 11, 12]. □

Thus, the total cost of connecting each g_i to some u_j is at most $2R_{\text{old}} \cdot n(R) \leq 2rR \cdot n(R)$. The expected cost of the tree, therefore, can be seen as bounded by the integral

$$\int 2rR \cdot n(R) d\mu.$$

Here, μ is the probability measure of the algorithm using balls of radius $2R$. Now, recall that once the initial value R_0 for R is chosen, we know that all balls used in the algorithm will have radius $2R_0/r^i$ for some integer i. Note that R_0 is a random variable given by $(\Delta/2)r^{-t}$, where t is chosen uniformly in $[0, 1]$. By inverting the expression for R_0 as a function to t, note that the probability that R_0 lies in a small range $[x, x + dx]$ is

$$\frac{\ln(x + dx) - \ln(x)}{\ln(r)} = \frac{\ln(1 + dx/x)}{\ln(r)} = \frac{dx}{x \ln(r)}.$$

Thus, the integral above is

$$\int_0^{\Delta/2} \frac{2r}{\ln(r)} \cdot n(R) dR.$$

By our lower bound, the algorithm produces a tree that is at most $2r/\ln(r)$ times the optimal cost. A simple calculation shows that this is minimized when $r = e$, and hence we have the following theorem.

Theorem 4.2. *The above algorithm achieves an expected approximation ratio of $2e$.*

The basic randomization technique we employed in the algorithm for choosing R_0 has been used previously in [9, 18].

Relative costs of minimum Steiner trees and zero skew trees. Using the lower bound we developed above, one can show that for n equally spaced terminals on a line, the optimal zero skew tree has $\Theta(\log n)$ times the cost of the minimum Steiner tree. We now give a short proof that this is, asymptotically, the largest possible gap. For purposes of simplicity, we do not attempt to optimize the constants in the proof.

Theorem 4.3. *For a set V of n points in a metric space M, let $\mathrm{St}(V)$ denote the cost of the minimum Steiner tree for V, and let $\mathrm{ZST}(V)$ denote the minimum cost of a zero skew tree for V. Then $\mathrm{ZST}(V) \leq (4e \ln n + e) \cdot \mathrm{St}(V)$.*

Proof. Recall that for a number R, $n(R)$ denotes the minimum number of balls of radius R needed to cover the terminals in V. We claim that for any R, $n(R) \leq 2\mathrm{St}(V)/R$.

To prove the claim, begin with an optimal Steiner tree and convert it into a traveling salesperson tour $\{v_{i_1}, \ldots, v_{i_n}\}$ of length at most $2\mathrm{St}(V)$, by doubling the edges and finding an Eulerian tour. We now construct a set $S \subseteq V$ as follows. We initially include v_{i_1}. Proceeding inductively, suppose S currently consists of $\{v_{i_1}, \ldots, v_{i_p}\}$. Let q denote the minimum index greater than p for which the length of the subtour from v_{i_p} to v_{i_q} is strictly greater than R; we add v_{i_q} to S and continue. At the end of this procedure, we observe that $|S| \leq 2\mathrm{St}(V)/R$, since the distance along the traveling salesperson tour between consecutive elements of S is at least R. Moreover, an arbitrary element $v_{i_j} \in V$ is within distance R of v_{i_t}, where t is the maximum index less than or equal to j for which $v_{i_t} \in S$; thus if we center a ball of radius R at each element of S, the resulting collection of balls covers V. The claim follows, since $n(R) \leq |S| \leq 2\mathrm{St}(V)/R$.

Applying the claim, and using the fact that $\mathrm{St}(V) \geq \Delta$, we have

$$\begin{aligned}
\mathrm{ZST}(V) &\leq 2e \int_0^{\Delta/2} n(R)dR \\
&= 2e \int_0^{\Delta/2n} n(R)dR + 2e \int_{\Delta/2n}^{\Delta/2} n(R)dR \\
&\leq 2e\left(\frac{\Delta}{2n}\right)n + 4e \int_{\Delta/2n}^{\Delta/2} \frac{\mathrm{St}(V)}{R} dR \\
&\leq e\Delta + 4e \cdot \mathrm{St}(V) \int_{\Delta/2n}^{\Delta/2} \frac{dR}{R} \\
&= e\Delta + 4e \cdot \mathrm{St}(V) \cdot \ln n \\
&\leq (4e \ln n + e) \cdot \mathrm{St}(V). \qquad \square
\end{aligned}$$

Derandomizing the algorithm. We briefly explain how the algorithm can be derandomized. Note that the only randomization used by the algorithm is in the initial choice of t, while setting $R_0 = (\Delta/2) \cdot \exp(-t \ln r)$. In the description of Algorithm Connect-Centers, the step where g_i is picked from S may be implemented arbitrarily; assume that this is implemented by some arbitrary, but fixed rule. We claim that the algorithm produces at most $O(n^2)$ combinatorially distinct trees and each of these can be produced by running the algorithm for $O(n^2)$ carefully chosen values of t. In order to see this, consider the kth iteration of Algorithm Connect-Centers (where iterations are numbered from 0 onwards). In this iteration, the value of $R = (\Delta/2) \cdot \exp(-(t+k)\ln r)$. Note that the choices of the algorithm depend only on which distances are at most $2R = \Delta \cdot \exp(-(t+k)\ln r)$. If for two values of t, the set of distances that are less than $2R$ is the same for all iterations, then the trees produced must have the same structure (since the choices made by the algorithm are exactly the same). Consider the distance between two vertices u and v. There is a unique value $t_{uv} \in [0,1)$ such that $d(u,v) = \Delta \cdot \exp(-(t_{uv}+k)\ln r)$, where k is an integer. Consider the set T of t_{uv} values for every pair of vertices u, v. (We also add 0 and 1 to T.) T can be easily determined and has $O(n^2)$ values. Further, for any t strictly between any two consecutive values (t_1, t_2) in T, the structure of the tree produced by the algorithm is exactly the same. It follows that the structure of the tree can be determined by running the algorithm for any $t \in (t_1, t_2)$. The edge costs can be set to the lowest possible value for t in this range, i.e., by pretending that we ran the algorithm for $t = t_2$. If we run this procedure for every pair of consecutive values in T, the best tree produced is at least as good as the expected cost of the randomized algorithm.

5. Bounded skew clock routing. We now present a constant-factor approximation algorithm for the bounded skew clock routing problem. The algorithm proceeds in two phases. First, we construct a Steiner tree spanning V which we fragment into subtrees. Second, we connect these subtrees using a modification of Algorithm Connect-Centers.

We first construct a Steiner tree T' spanning V. To do this, we use the currently best known approximation algorithm for Steiner trees in general metric spaces due to [20]. In case the terminals are in the plane, we can use a polynomial time approximation scheme (PTAS) for Steiner trees [1], with an approximation ratio of $1+\epsilon$ for any $\epsilon > 0$. Let $W \subseteq V$ be a maximal subset of terminals such that the distance between any two of them in T' is at least s. W can be chosen by a greedy algorithm.

LEMMA 5.1.

$$|W| \leq 2\text{cost}(T')/s.$$

Proof. For $v \in W$, let B_v be a ball of radius $s/2$ about v, distances being computed in the metric induced by the tree T'. Then, for $u, v \in W$, $B_u \cap B_v = \emptyset$. Now, the Steiner tree T' has a path P_v of length $s/2$ within each ball B_v. (Here, P_v could include a fractional part of an edge.) The sum of the lengths of the paths P_v is at most cost(T'). Hence, the number of paths (and therefore, the number of vertices in W) is at most $2\text{cost}(T')/s$. □

For each $v \in W$, we construct a tree T_v rooted at v, such that the distance from v to every vertex in T_v is at most s. To do this, we order the vertices in W arbitrarily, say $W = \{v_1, \ldots, v_k\}$. Now, we assign every vertex in V to the closest vertex in W, breaking ties in favor of vertices with smaller indices. Here distances are computed in T'. Note that every vertex in V is within a distance of at most s from some vertex in W (by the maximality of W). For $v \in V$, let $c(v)$ denote the vertex in W that it is assigned to; let P_v denote the path in T' from v to $c(v)$. The length of P_v is at most s.

LEMMA 5.2. *For $v_1 \neq v_2$, if $c(v_1) \neq c(v_2)$, the paths P_{v_1} and P_{v_2} are edge disjoint.*

Proof. Suppose P_{v_1} and P_{v_2} share an edge. An easy case analysis shows that this contradicts the choice of either $c(v_1)$ or $c(v_2)$. □

For vertex $v \in W$, let $S(v)$ denote the set of vertices assigned to it. Let T_v be the subtree of T' that spans $S(v)$; in other words, $T_v = \cup_{u \in S(v)} P_u$. Then Lemma 5.2 implies that the subtrees T_v are disjoint. Clearly, we also have the following lemma.

LEMMA 5.3.

$$\sum_{v \in W} \text{cost}(T_v) \leq \text{cost}(T').$$

Also, the distance from v to every vertex in T_v is at most s.

Now we describe how to modify Algorithm Connect-Centers using the subtrees T_v constructed above to produce the final tree with skew at most s. We execute Algorithm Connect-Centers, but stop the process of construction of the tree at the last step when $R < s$ for the first time (i.e., we stop *before* a value for R smaller than s is used to create a partition). Let R_f be the final value of R (so $R_f < s$). At this time, $\bar{U}$ is a partition of V such that every vertex in U_i is at a distance at most $2rR_f$ from u_i. Let T be the partial tree constructed by the algorithm so far. We will connect each of the subtrees T_v to the tree T in the following way: For $v \in W$, let i_v be such that $v \in U_{i_v}$. Connect T_v to the tree T by adding an edge of weight $2rR_f$ from v to u_{i_v}. It is easy to see that the tree so constructed has skew at most s.

Now, we shall analyze the cost of the tree we obtain. Let C_1 be the cost of the tree that the algorithm constructs until $R < s$ for the first time. Let C_2 be the total cost of all the edges from vertices $v \in W$ to u_{i_v}. Let C_3 be the total cost of the trees T_v for $v \in W$.

Then, by the previous analysis,

$$\mathbf{E}[C_1] \leq \frac{2r}{\ln r} \int_s^{\Delta/2} n(R) dR.$$

Also,

$$C_3 \leq \operatorname{cost}(T').$$

Now,

$$\begin{aligned}\mathbf{E}[2rR_f] &= 2\int_s^{rs} \frac{1}{\ln r} dx \\ &= \frac{2(r-1)s}{\ln r}.\end{aligned}$$

Hence,

$$\begin{aligned}\mathbf{E}[C_2] &= |W| \cdot \mathbf{E}[2rR_f] \\ &= |W| \frac{2(r-1)s}{\ln r} \\ &\leq \frac{4(r-1)}{\ln r} \operatorname{cost}(T').\end{aligned}$$

Let OPT_{ST} be the cost of the optimal Steiner tree on the set of terminals. Since the Steiner tree T' is constructed using the algorithm of [20], this guarantees that

$$\operatorname{cost}(T') \leq \left(1 + \frac{\ln 3}{2}\right) \mathsf{OPT}_{ST}.$$

Let OPT be the cost of the optimal clock tree with skew at most s. Then, we have two lower bounds for OPT. Using the lower bound given in section 3,

$$\begin{aligned}\mathsf{OPT} &\geq \int_s^{\Delta/2} n(R) dR. \\ \mathsf{OPT} &\geq \mathsf{OPT}_{ST}.\end{aligned}$$

Now, we can bound the expected cost of the tree we obtain in terms of OPT as follows:

$$\mathbf{E}[C_1 + C_2 + C_3] \leq \left(\frac{2r}{\ln r} + \left(1 + \frac{\ln 3}{2}\right)\left(1 + \frac{4(r-1)}{\ln r}\right)\right) \mathsf{OPT}.$$

The approximation ratio is optimized by choosing $r \approx 1.775$, which gives an approximation ratio of at most 16.1065.

6. Planar zero skew clock routing.

We now present a constant-factor approximation algorithm for the planar ZST problem. In the planar case we refer to vertices as points.

Let R^* be the smallest radius of an L_1 ball that encloses all the terminals. We first find a center point r such that every terminal is within an L_1 distance of R^* from r. We now construct a square S of side $2R_0$ centered at r. The value R_0 is chosen by selecting t uniformly and at random from $[0, 1]$ and setting $R_0 = R^* \cdot 2^t$. The square

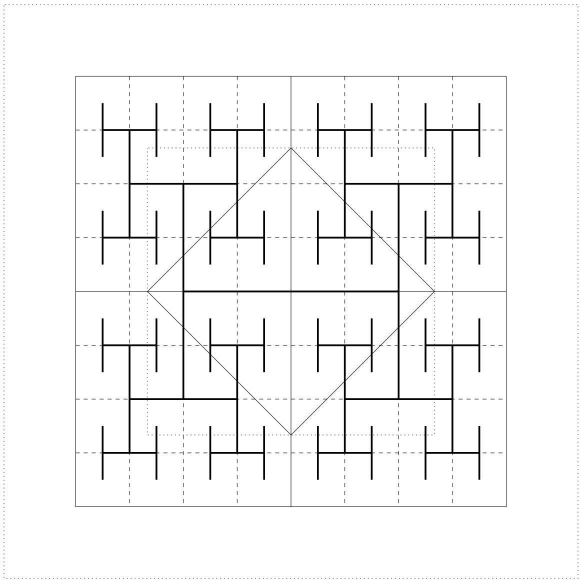


FIG. 2. *The first few levels of an H-tree.*

S is then subdivided into four equal-sized squares S_1, S_2, S_3, S_4. The squares S_i are called the *children* of S, and S is called the *parent* of each S_i. The center of S is connected to the centers of each S_i by an H-shaped structure. We proceed recursively in each S_i, dividing each into four equal squares and so on, so long as there is at least one point in the square. This produces a tree that we refer to as an *H-tree* (see Figure 2.)

This tree spans all the terminals. In fact, we construct only the subtree of the H-tree that spans all the terminals. To do this, we ensure that the tree construction proceeds only inside squares that contain at least one terminal. At any point in the execution of the algorithm, consider a square S produced by the algorithm and subdivided into S_1, S_2, S_3, S_4. Then the center of S is connected to the center of S_i, and the tree construction proceeds recursively in S_i only if S_i contains a terminal. Also, we stop the recursive subdivision when the squares that the algorithm constructs have side lengths smaller than R^*/n^2. At this stage, we connect the centers of all squares to the terminals inside them by edges of length R^*/n^2.

In order to analyze the cost of the tree returned by the algorithm, we associate, with each square S constructed by the algorithm, the cost of the connection from the center of S to the center of the parent of S. Thus, the charge to a square of side $2x$ is $2x$. Note that when the algorithm terminates, the cost of connecting the n terminals to the centers of their corresponding squares is $n \cdot R^*/n^2 = R^*/n$. Since the cost of the optimal ZST is at least $2R^*$, this is at most $1/n$ times the optimal cost. We ignore this cost in our calculations, and in fact, the algorithm can be modified so that this cost is not incurred.

Let $n(x)$ be the minimum number of L_1 balls of radius x required to cover the terminals. Let $n'(y)$ be the number of squares of radius y produced by the algorithm.

LEMMA 6.1.

$$n'(2x) \leq 4n(x).$$

Proof. Consider a grid with the center point r as the origin, produced by equispaced horizontal and vertical lines such that the distance between consecutive lines is $2x$. The squares of side $2x$ produced by the algorithm are precisely the squares in this grid that contain at least one terminal. Consider the optimal partitioning of the terminals into $n(x)$ L_1 balls of radius x. Each L_1 ball in the partition intersects at most four squares in the grid. Thus there can be at most $4n(x)$ squares in the grid that contain at least one terminal. □

Since the algorithm constructs squares of side lengths in the range $[0, 2R^*]$, the expected cost of the tree is bounded by

$$\int_0^{2R^*} n'(y) \cdot y \cdot d\mu.$$

Here, $d\mu$ is the probability that the algorithm constructs squares with side length in the range $[y, y + dy]$. Hence $d\mu = dy/(y \ln 2)$. The expected cost is thus bounded by

$$\begin{aligned}\int_0^{2R^*} \frac{n'(y)}{\ln 2} dy &= \int_0^{R^*} \frac{n'(2x)}{\ln 2} 2dx \\ &\leq \int_0^{R^*} \frac{8}{\ln 2} n(x) dx.\end{aligned}$$

Hence, the expected cost of the tree produced by the algorithm is at most $8/\ln 2 \approx 11.54$ times the optimal cost.

THEOREM 6.2. *The above algorithm achieves an expected approximation ratio of* $8/\ln 2$.

The algorithm can be derandomized easily by carefully choosing a set of $O(n^2)$ values of R_0 in the range $[R^*, 2R^*]$, running the algorithm for each of them, and returning the best tree produced. The details are similar to those in the derandomization of the algorithm in section 4.

7. Planar-embeddable bounded skew clock routing. We now give a constant factor approximation algorithm for creating planar-embedded s-skew trees. We apply the lower bound of section 3, namely, $\mathsf{OPT} \geq \int_s^{R^*} n(R) dR$.

Our strategy will be similar to the bounded skew case for general metrics. We construct the zero skew tree as in the previous section but stop when the sides of the squares become smaller than $s/2$. We will then connect the points in each square to the center using a tree whose cost is comparable to the minimum spanning tree (MST) for the point set and has radius at most s. We will separately bound the cost of both the truncated ZST and the trees within each square to within a constant factor of OPT.

We present a deterministic version of the algorithm here. Let R_0 be the unique value in $[R^*, 2R^*)$ of the form $2^t s$, where t is integral. Let $R_i = 2^{t-i}s$. Enclose the point set in a box of side $2R_0$. We iteratively divide the square as before into four squares, but we stop after $t+1$ iterations, when the side of the resulting square has size $s/2$. We then build a zero skew H-tree terminating at the centers of every populated square of size $s/2$. We will now connect the points within each square to the center.

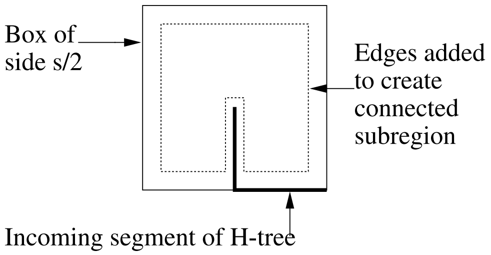

Box of side s/2
Edges added to create connected subregion
Incoming segment of H-tree

Fig. 3. *New edges added to each box.*

We first construct an MST connecting all the points in the point set. We divide this MST into pieces using the ZST built above. Recall that the ZST includes a single edge into the center of each square. We cut each edge in the MST at points where it intersects existing edges in the ZST, or boundaries of the squares of side $s/2$. We augment the MST edges within each square to produce a connected planar graph, by adding the new edges shown in Figure 3. This results in a connected graph within each square of side $s/2$, from which we take any spanning subtree.

We apply the following result (see [2, 16]).

Lemma 7.1. *Given any $\epsilon > 0$ and point set in the plane with radius r, and spanning tree T with cost c rooted at p, there exists a polynomial time algorithm to find a spanning tree T' with radius $r' \leq (1+\epsilon)r$ and cost $c' \leq (1+1/\epsilon)c$.*

We run this algorithm for $\epsilon = 1$ on each square of side $s/2$, and attach the resulting spanning tree to the ZST at the center of the square.

Now, notice that the cost of the resulting structure has two components, each of which we bound separately. First, the edges of the ZST and the additional edges of Figure 3 are bounded by five times the cost of the ZST. We can bound the cost of the ZST using techniques similar to those of the previous section, with the caveat that rather than bounding $n'(x)$ in terms of $n(x/2)$, we instead bound it in terms of $n(x/4)$.

8. Hardness of zero skew clock routing.

Theorem 8.1. *The zero skew clock routing problem for general metric spaces is* **NP***-hard.*

Proof. The reduction is from set cover. Let $[n] = \{1, 2, \ldots, n\}$. A set cover instance consists of an integer k, and m sets $S_1, \ldots, S_m$ such that each S_i is a subset of $[n]$. We are required to determine if there exist k sets $S_{i_1}, \ldots, S_{i_k}$ such that $[n] \subseteq \bigcup_{j=1}^{k} S_{i_j}$. Given an instance I of set cover, we define an instance of the zero skew clock routing problem as follows. We first construct a weighted graph G from the instance I: G has a vertex s, vertices $x_1, \ldots, x_m$ (one corresponding to each set), and vertices $y_1, \ldots, y_n$ (one corresponding to each element of $[n]$) and an edge from x_i to y_j iff $j \in S_i$; such an edge has length 1. Also, every x_i is connected to s by an edge of length $1/n$. Consider the zero skew clock routing problem for the set of terminals $\{y_1, \ldots, y_n\}$ in the metric space induced by distances in G. If k' is the minimum number of sets in the instance I required to cover $[n]$, it is easy to show that the optimal solution to the zero skew clock routing problem is $n + k'/n$. □

9. Open questions. The complexity of the planar ZST problem is still open. We do not know if the problem is **NP**-hard.

Since our algorithm can be thought of yielding a clock tree topology, it will be interesting to see how it performs in practice, especially when combined with the DME technique.

REFERENCES

[1] S. Arora, *Polynomial time approximation schemes for Euclidean traveling salesman and other geometric problems*, J. ACM, 45 (1998), pp. 753–782.

[2] B. Awerbuch, A. Baratz, and D. Peleg, *Cost-sensitive analysis of communication protocols*, in Proceedings of the 9th Annual ACM Symposium on Principles of Distributed Computing, New York, 1990, pp. 177–187.

[3] H. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison–Wesley, Reading, MA, 1990.

[4] T.-H. Chao, Y.-C. Hsu, J.-M. Ho, K. D. Boese, and A. B. Kahng, *Zero skew clock routing with minimum wirelength*, IEEE Trans. Circuits and Systems Part 2: Analog and Digital Signal Processing, 39 (1992), pp. 799–814.

[5] J. Cong, A. Kahng, and G. Robins, *Matching-based methods for high-performance clock routing*, IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, 12 (1993), pp. 1157–1169.

[6] J. Cong, A. Kahng, C.-K. Koh, and C.-W. Tsao, *Bounded-skew clock and Steiner routing under Elmore delay*, TR-950030, Computer Science Department, University of California, Los Angeles, 1995.

[7] J. Cong and C.-K. Koh, *Minimum-cost bounded-skew clock routing*, TR-950003, University of California, Los Angeles, Computer Science Department, 1995.

[8] M. Edahiro, *An efficient zero-skew routing algorithm*, in Proceedings of the 31st ACM IEEE Design Automation Conference, ACM, New York, 1994, pp. 375–380.

[9] M. Goemans and J. Kleinberg, *An improved approximation ratio for the minimum latency problem*, in Proceedings of the 7th Annual ACM–SIAM Symposium on Discrete Algorithms, 1996, pp. 152–157.

[10] T. E. Gonzalez, *Clustering to minimize the maximum intercluster distance*, Theoret. Comput. Sci., 38 (1985), pp. 293–306.

[11] D. Hochbaum, *Various notions of approximations: Good, better, best, and more*, in Approximation Algorithms for NP-Hard Problems, D. Hochbaum, ed., PWS Publishing Company, Boston, MA, 1996, pp. 346–398.

[12] D. S. Hochbaum and D. B. Shmoys, *A best possible heuristic for the k-center problem*, Math. Oper. Res., 10 (1985), pp. 180–184.

[13] M. Jackson, A. Srinivasan, and E. Kuh, *Clock routing for high-performance ICs*, in Proceedings of the 27th ACM IEEE Design Automation Conference, ACM, New York, 1991, pp. 573–579.

[14] A. Kahng and G. Robins, *On Optimal Interconnections for VLSI*, Kluwer Academic Publishers, Norwell, MA, 1995.

[15] A. Kahng, J. Cong, and G. Robins, *High-performance clock routing based on recursive geometric matching*, in Proceedings of the 28th ACM IEEE Design Automation Conference, ACM, New York, 1991, pp. 322–327.

[16] S. Khuller, B. Raghavachari, and N. Young, *Balancing minimum spanning trees and shortest-path trees*, Algorithmica, 14 (1995), pp. 305–321.

[17] Y.-M. Li and M. Jabri, *A zero-skew clock routing scheme for VLSI circuits*, in Proceedings of the IEEE International Conference on Computer-Aided Design, IEEE Computer Society, Los Alamitos, CA, 1992, pp. 458–463.

[18] R. Motwani, S. Phillips, and E. Torng, *Non-clairvoyant scheduling*, Theoret. Comput. Sci., 130 (1994), pp. 17–47.

[19] P. Ramanathan, A. J. Dupont, and K. G. Shin, *Clock distribution in general VLSI circuits*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 41 (1994), pp. 395–404.

[20] G. Robins and A. Zelikovsky, *Improved Steiner tree approximation in graphs*, in Proceedings of the 11th Annual ACM–SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 2000, pp. 770–779.

[21] R.-S. Tsay, *Exact zero skew*, in Proceedings of the IEEE International Conference on Computer-Aided Design, IEEE Computer Society, Los Alamitos, CA, 1991, pp. 336–339.

[22] R.-S. Tsay, *An exact zero-skew clock routing algorithm*, IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, 12 (1993), pp. 242–249.

[23] J. G. Xi and W. W.-M. Dai, *Jitter-tolerant clock routing in two-phase synchronous systems*, in Proceedings of the IEEE International Conference on Computer-Aided Design, IEEE

Computer Society, Los Alamitos, CA, 1996, pp. 316–321.

[24] A. Z. Zelikovsky and I. I. Măndoiu, *Practical approximation algorithms for zero- and bounded-skew trees*, SIAM J. Discrete Math., 15 (2002), pp. 97–111.

[25] Q. Zhu and W. W.-M. Dai, *Perfect-balance planar clock routing with minimal path-length*, in Proceedings of the IEEE International Conference on Computer-Aided Design, IEEE Computer Society, Los Alamitos, CA, 1992, pp. 473–477.

SIAM J. Discrete Math.
Vol. 17, No. 4, pp. 596–614

ENCODING FULLERENES AND GEODESIC DOMES[*]

JACK E. GRAVER[†]

Abstract. Coxeter's classification of the highly symmetric geodesic domes (and, by duality, the highly symmetric fullerenes) is extended to a classification scheme for all geodesic domes and fullerenes. Each geodesic dome is characterized by its signature: a plane graph on twelve vertices with labeled angles and edges. In the case of the Coxeter geodesic domes, the plane graph is the icosahedron, all angles are labeled one, and all edges are labeled by the same pair of integers (p, q). Edges with these "Coxeter coordinates" correspond to straight line segments joining two vertices of Λ, the regular triangular tessellation of the plane, and the faces of the icosahedron are filled in with equilateral triangles from Λ whose sides have coordinates (p, q).

We describe the construction of the signature for any geodesic dome. In turn, we describe how each geodesic dome may be reconstructed from its signature: the angle and edge labels around each face of the signature identify that face with a polygonal region of Λ and, when the faces are filled by the corresponding regions, the geodesic dome is reconstituted. The signature of a fullerene is the signature of its dual. For each fullerene, the separation of its pentagons, the numbers of its vertices, faces, and edges, and its symmetry structure are easily computed directly from its signature. Also, it is easy to identify nanotubes by their signatures.

Key words. fullerenes, geodesic domes, nanotubes

AMS subject classifications. 05C10, 92E10, 52A25

DOI. 10.1137/S0895480101391041

1. Introduction.

By a fullerene, we mean a trivalent plane graph $\Phi = (V, E, F)$ with only hexagonal and pentagonal faces. It follows easily from Euler's formula that each fullerene has exactly 12 pentagonal faces. In this paper, we work with the duals to the fullerenes: geodesic domes, i.e., triangulations of the sphere with vertices of valence 5 and 6. It is in this context that Coxeter [3], Caspar and Klug [2], and Goldberg [7] parameterized the geodesic domes/fullerenes that include the full rotational group of the icosahedron among their symmetries. These highly symmetric geodesic domes are obtained by filling in each face of the icosahedron with a fixed equilateral triangle inscribed in Λ, the regular triangular tessellation of the plane. Coxeter's classification boils down to classifying the equilateral triangles of Λ.

Our plan is to extend Coxeter's approach to other plane graphs with 12 vertices, filling in their faces with regions from Λ such that the original 12 vertices are the vertices of valence 5 in the resulting geodesic dome. These special planar graphs with 12 vertices will be called signature graphs. The signature graph along with the labeling of the edges and angles that determines just how the faces are to be filled in will be called the signature of the resulting geodesic dome.

Let $\Phi = (V, E)$ be any graph with a set of edge weights, $\omega : E \to \mathbb{R}^+$. The *structure graph* of the weighted graph Φ, ω is the union of all shortest spanning trees of Φ. Now let $\Gamma = (V, E, F)$ be a geodesic dome and let P denote the set of the 12, 5-valent vertices of Γ. The first step in constructing the signature graph of Γ is to construct the complete graph on the vertex set P and assign to each of its edges the distance between its endpoints, as vertices in Γ. This weighted graph is called the *first auxiliary graph of* Γ and is denoted by $\mathcal{A}_1(\Gamma)$. The second step is to construct

[*] Received by the editors June 20, 2001; accepted for publication (in revised form) September 8, 2003; published electronically May 17, 2004.
http://www.siam.org/journals/sidma/17-4/39104.html

[†] Department of Mathematics, Syracuse University, Syracuse, NY 13244-1150 (jegraver@syr.edu).

the structure graph of $\mathcal{A}_1(\Gamma)$. This graph is called the *second auxiliary graph of* Γ and is denoted by $\mathcal{A}_2(\Gamma)$. This graph, $\mathcal{A}_2(\Gamma)$, has a natural drawing on the sphere but may admit crossing edges. To eliminate any crossings, we make a slight alteration in the weight function and construct the structure graph of $\mathcal{A}_2(\Gamma)$ with this new weight function to get a third graph. This third graph also has a natural drawing on the sphere and it admits no crossings. The plane graph consisting of this third graph and its natural planar embedding is called the *signature graph* of Γ and is denoted by $\mathcal{S}(\Gamma)$. Each edge of $\mathcal{S}(\Gamma)$ may be identified with a line segment joining two vertices in Λ. This identification leads to a labeling system for the edges and angles of $\mathcal{S}(\Gamma)$. The signature graph of Γ along with this labeling is called the *signature* of Γ. We should note that Coxeter's approach has been generalized to some other triangulations of the sphere by Fowler, Cremona, and Steer [4] and by Fowler and Cremona [5] using an entirely different labeling system.

Each geodesic dome Γ is completely determined by its signature. Using the signature of Γ as a blueprint, one can construct a polygonal region or a set of polygonal regions in Λ with sides corresponding to the edges of $\mathcal{S}(\Gamma)$ and then glue them together to reconstruct Γ. Since all signature graphs of geodesic domes have exactly 12 vertices and since any planar graph admits only a finite number of distinct planar embeddings, there are only a finite number of plane graphs which could be the signature graph of a geodesic dome. This leads to a partition of the collection of all geodesic domes into a finite number of classes each corresponding to a different signature graph. We may label the angles of a given signature graph in a finite number of ways and we may label the edges with variables in a finite number of ways. Hence, within each class, we have a finite number of families. Each family corresponds to a signature graph with labeled angles and with variable labels on the edges. Each choice of the variables, satisfying an included set of equalities and inequalities, will then yield the signature of a specific geodesic dome or fullerene. The geodesic domes described by Coxeter form such a family.

We develop the signature in a more general setting defining it for each *plane triangulation*, that is for each plane graph with only triangular faces. To carry out these tasks, we will need several tools. We start our investigation with a short section on the basic properties of structure graphs followed by an extensive development of the "geometry" of Λ.

2. Structure graphs.

LEMMA 1. *Let Θ be the structure graph of the weighted graph Φ, ω.*

i. *If u, v, w are vertices of a 3-circuit in Φ with $\omega(\{u,w\}) < \omega(\{v,w\})$ and $\omega(\{u,v\}) < \omega(\{v,w\})$, then the edge $\{v,w\}$ is not in Θ.*

ii. *Deleting the edges of maximum weight from Θ disconnects Θ.*

iii. *If the edges of Θ of maximum weight are deleted from Θ, then each of the resulting components is the structure graph of the corresponding vertex induced weighted subgraph of Φ.*

iv. *If Ω is any connected subgraph of Θ, then deleting the edges of maximum weight from Ω disconnects Ω.*

Proof. Let $\Phi = (V,E)$ and $\Theta = (V,F)$.

Part i. Suppose that the edge $e = \{v,w\}$ is in Θ and let (V,T) be a shortest spanning tree of Φ containing e. Delete e from (V,T). The vertex u is either in the component of $(V,T-e)$ which contains v or in the component which contains w. If it is in the component containing v, then $(V,T-e+e')$, where $e' = \{u,w\}$, is a shorter spanning tree; if u is in the component containing w, then $(V,T-e+e'')$, where

$e'' = \{u, v\}$, is a shorter spanning tree. Since both possibilities contradict the fact that (V, T) is a shortest spanning tree, our supposition must be false.

Part ii. Let m denote the maximum among the weights of the edges of Θ and let $e = \{v, w\}$ be an edge of Θ with weight m. Let (V, T) be a shortest spanning tree of Φ which contains e. Delete e from (V, T). We show that each edge in the cutset of the edges of Θ joining the two components of $(V, T - e)$ has weight m. Let e' be any edge of Θ with an endpoint in each of the two components. Thus $(V, T - e + e')$ is also a spanning tree of Φ. This new tree has weight $\omega(T) - m + \omega(e')$, where $\omega(T)$ denotes the sum of the weights of the edges in T. Since (V, T) is a shortest spanning tree $\omega(T) \leq \omega(T) - m + \omega(e')$ or $m \leq \omega(e')$. But, $\omega(e') \leq m$; hence e' has weight m.

Part iii. Let m denote the maximum weight of the edges in Θ and let U be the vertex set of a component of the subgraph of Θ obtained by deleting all edges of weight m. Let $\Phi' = (U, G)$ and $\Theta' = (U, H)$ be the subgraphs of Φ and Θ induced by U. We show that Θ' is the structure graph of Φ'. Let (V, T) be a shortest spanning tree of Φ and let (U, T') be the subgraph of this tree induced by U. We assert that (U, T') is a shortest spanning tree of Φ'.

Suppose that (U, T') is not connected. If e is any edge of Θ' joining two components of (U, T') and e' is any edge in $T - T'$ incident to one of these components, we have that $\omega(e) < m = \omega(e')$ and that $(V, T - e' + e)$ is a shorter spanning tree of Φ. Thus (U, T') is a spanning tree of Φ'. Let (U, T'') be any shortest spanning tree of Φ'. Then $(V, T - T' + T'')$ is a spanning tree of Φ and $\omega(T'') \leq \omega(T')$. It follows that $\omega(T'') = \omega(T')$, that (U, T') is a shortest spanning tree of Φ' and that $(V, T - T' + T'')$ is a shortest spanning tree of Φ. Thus Θ' is the union of the shortest spanning trees of Φ'.

Part iv. We proceed by induction on the number of vertices of Φ. Let Ω be any connected subgraph of Θ. If the maximum weight of the edges in Ω equals the maximum weight of the edges in Θ, then, by part ii, deleting the edges of this maximum weight disconnects Ω. If the maximum weight of the edges in Ω is less than the maximum weight of the edges in Θ, then Ω is a subgraph of the structure graph of the smaller weighted graph Φ' induced by the vertex set of the component containing Ω of the graph obtained by deleting all edges of maximum weight from Θ. We may now apply the induction hypothesis. □

3. The regular triangular tessellation of the plane.

Consider Λ, the regular triangular tessellation of the plane. We think of Λ as the infinite plane graph with all vertex valences 6 and all face valences 3. The automorphisms of this graph correspond with the congruences of Λ as a geometric object in the plane: the translations, rotations, reflections, and glide reflections that map Λ onto Λ. Two vertex sets of Λ are said to be *congruent* if there is an automorphism of Λ which maps one onto the other. By a *segment* of Λ we simply mean a pair of vertices of Λ and we visualize a segment as the straight line segment joining the two vertices. To each segment which does not coincide with a "line" of the tessellation, we assign *Coxeter coordinates* (p, q) as follows: select one endpoint of the segment to be the origin; take the edge of the graph to the right of the segment as the unit vector in the p direction; take the edge of the graph to the left of the segment as the unit vector in the q direction; finally, assign to the segment the coordinates of its other endpoint in this coordinate system. If the segment coincides with a "line" of the tessellation, that segment is assigned the single Coxeter coordinate (p), where p is the number of edges of Λ in the segment. In Figure 1, we illustrate this definition by giving the Coxeter coordinates assigned to the sides of several different regions in Λ. The *length* of a segment σ with endpoints

v and w is defined to be the graph-theoretic distance between the endpoints in the graph Λ; it is denoted by $\delta(v, w)$ or $|\sigma|$. Collected in the next lemma are several observations about this labeling of segments. The proofs are straightforward.

LEMMA 2. *Let σ denote a segment with endpoints v and w and let (p, q) [or (p)] denote its Coxeter coordinates as computed from v.*

i. *The Coxeter coordinates of σ computed from w are also (p, q) [(p)].*
ii. *p and q are positive integers (p is a positive integer).*
iii. *$|\sigma| = p + q$ ($|\sigma| = p$).*
iv. *The segments σ, with Coxeter coordinates (p, q), and σ', with Coxeter coordinates (p', q'), are congruent if and only if either $p' = p$ and $q' = q$ or $p' = q$ and $q' = p$. Furthermore, $p' = p$ and $q' = q$ if and only if σ' is the image of σ under a rotation or translation of the tessellation and $p' = q$ and $q' = p$ if and only if σ' is the image of σ under a reflection or glide reflection of the tessellation.*
v. *The segments σ, with Coxeter coordinate (p), and σ', with Coxeter coordinate (p'), are congruent if and only if $p' = p$. Furthermore, any two segments with coordinates (p) are images of one another under both a translation or rotation and a reflection or glide reflection.*

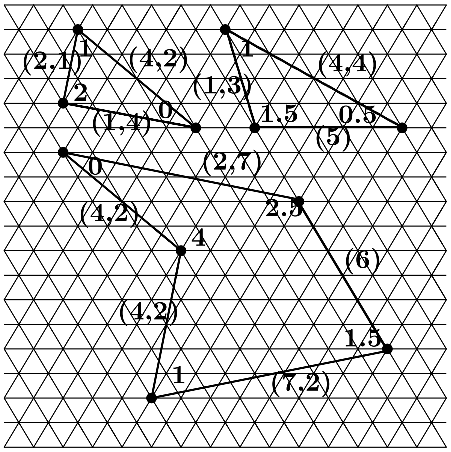


FIG. 1.

We are particularly interested in angles. Much of the information about an angle is coded in the Coxeter coordinates of its sides, but not all. Suppose that we have two segments forming an angle at a common endpoint v; denote them, in clockwise order, by σ and σ', denoting their Coxeter coordinates by (p, q) and (p', q'), respectively. The missing information is the multiple of 60 degrees between the edge from v along which p is measured and the edge along which p' is measured. This multiple is easily seen to be the number of edges from v which lie between the two segments. Hence, we define the *type* of the angle between two segments with a common endpoint v to be the number of edges from v which lie between the two segments. Segments with Coxeter coordinates of the form (p) coincide with an edge; in this case, the edge contributes $\frac{1}{2}$ to each of the types of the angles on either side of the segment. These definitions are illustrated in Figure 1 and the next lemma lists some useful properties of angle type.

LEMMA 3.

i. *Given segments α, β, and γ in clockwise order around a common endpoint,* *the type of the angle between α and γ is the sum of the types of the angles between α and β and between β and γ.*

ii. *Given segments $\sigma_1, \sigma_2, \ldots, \sigma_n$ with a common endpoint, the sum of the types of the angles between them is 6.*

iii. *Given an n-gon with angle types $A_1, A_2, \ldots, A_n$, we have*
$A_1 + \cdots + A_n = 3n - 6$.

Proof. Part i follows directly from the definition of angle type, and part ii follows directly from part i.

Turning to part iii, consider a triangle with vertices labeled v_A, v_B, v_C in clockwise order with corresponding angle types A, B, and C. Let α denote the segment opposite v_A; β, the segment opposite v_B; and γ, the segment opposite v_C. For any segment σ with Coxeter coordinates (p, q), let θ_σ denote the measure, in degrees, of the angle between the segment and the lattice edge in the direction along which p is measured; note that θ_σ is independent of the endpoint of σ at which it is measured. For a segment with Coxeter coordinate (p), define θ_σ to be 30°. We observe that, in degrees, the measure of the angle at v_A is $60A - \theta_\beta + \theta_\gamma$; see Figure 2. Similarly, the measure of the angle at v_B is $60B - \theta_\gamma + \theta_\alpha$ and the measure of the angle at v_C is $60C - \theta_\alpha + \theta_\beta$. Summing the measures of these three angles gives $60A + 60B + 60C = 180$ or $A + B + C = 3$.

Since each n-gon can be partitioned into $n - 2$ triangles, the result follows from this special case and part i. □

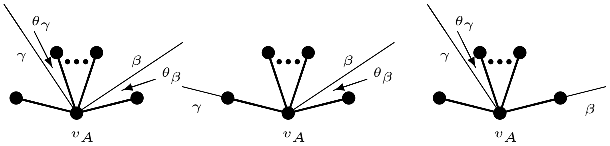


FIG. 2.

Our next task is to explore the structure of Λ in the neighborhood of a segment. We start by characterizing shortest paths in Λ and certain shortest paths in an arbitrary plane triangulation Γ. To do this, we introduce some additional terminology. Let $v = v_0, v_1, \ldots, v_n = w$ be any path from v to w in Λ or any v, w-path in Γ such that $v_1, \ldots, v_{n-1}$ are 6-valent. Consider the vertex v_i, $i = 1, \ldots, n-1$, and consider the vertices adjacent to v_i clockwise from v_{i-1}. If v_{i+1} is in the first position, we say the path takes a *sharp left turn* at v_i; v_{i+1} in the second position corresponds to a *left turn*; v_{i+1} in the fourth position corresponds to a *right turn*; v_{i+1} in the fifth position corresponds to a *sharp right turn*; otherwise we say the path continues *straight on* at v_i.

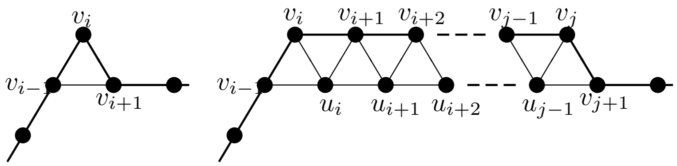


FIG. 3.

First, observe that if there is a sharp right (sharp left) turn at v_i, then $v = v_0, v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n = w$ is a shorter v, w-path. This is pictured at the left in Figure 3. Next, suppose that the path takes a right (left) turn at v_i, then continues straight on to v_j, where it takes another right (left) turn, pictured at the right in

Figure 3. Let u_i be the vertex adjacent to v_{i-1}, v_i, v_{i+1}; then let u_k be the vertex adjacent to u_{k-1}, v_k, v_{k+1}, for $k = i+1, \ldots, j-1$ (see Figure 3). Again we see that we have a shorter v, w-path:

$$v = v_0, v_1, \ldots, v_{i-1}, u_i, \ldots, u_{i-1}, v_{i+1}, \ldots, v_n = w.$$

Now suppose that we have a v, w-path Π in Λ which makes no sharp turns and in which the turns alternate between left and right. One easily verifies that this is a shortest v, w-path. We have the following lemma.

Lemma 4. *Let Π denote a shortest path in Λ or a shortest path in a plane triangulation Γ that has only 6-valent interior vertices. Then Π makes no sharp turns and consecutive turns alternate between left and right. Furthermore, a path in Λ which makes no sharp turns and in which the turns alternate between left and right is a shortest path between its endpoints.*

Consider a segment σ in Λ with endpoints v and w and Coxeter coordinates (p, q). From this lemma, we easily see that all shortest paths from v to w lie in the parallelogram with antipodal vertices v and w and with sides parallel to the p and q directions. Furthermore, all v, w-paths in this parallelogram using only edges in the p or q directions are shortest v, w-paths. We call these the *family of shortest paths associated with the segment* σ and denote this family of paths and the parallelogram containing them by G_σ.

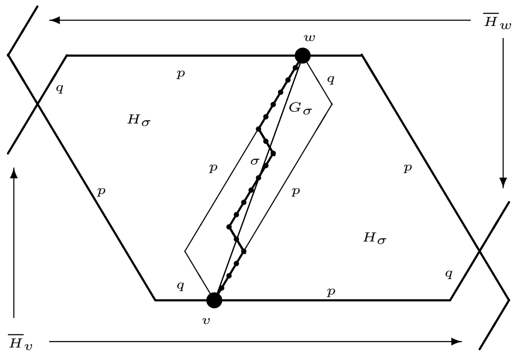


Fig. 4.

Now construct the hexagonal circuit $\overline{H}_v$ about v spanned by the vertices at a distance of $p + q$ from v and let H_v denote the finite subgraph of Λ bounded by $\overline{H}_v$. Define $\overline{H}_w$ and H_w similarly. Let $H_\sigma = H_v \cap H_w$ and denote its bounding circuit by $\overline{H}_\sigma$. We have pictured the various regions and boundaries of this construction in Figure 4. In drawing this picture, we have made some assumptions, namely that $0 < q < p$. If $q = p$, the picture is the same but with the segment vertical and, if $q > p$, the picture is the mirror image of this picture with the p and q labels reversed. $\overline{H}_\sigma$ is a hexagon with opposite sides parallel and equal in length. The points v and w are antipodal on this boundary dividing the sides containing them into segments of length p and q. If σ has Coxeter coordinate (p), H_σ consists of the union of two equilateral triangles with the given segment as the common side. In this case, $\overline{H}_\sigma$ is a rhombus with sides of length p and may be visualized by letting $q = 0$ in Figure 4. Some easily checked but useful properties of this configuration are listed in the next lemma.

LEMMA 5.

a. *Let the segment σ, with endpoints v and w and Coxeter coordinates (p, q), be given and let G_σ, H_σ, $\overline{H}_\sigma$, $\overline{H}_v$, and $\overline{H}_w$ be defined as above.*
   - (i) *The collection of shortest v, w-paths in Λ is the collection of all v, w-paths in the parallelogram G_σ using only edges in the p or q directions.*
   - (ii) *For any vertex $u \in H_\sigma$, $\delta(u, v) \leq |\sigma|$ and $\delta(u, w) \leq |\sigma|$ with strict inequality in both cases whenever $u \in H_\sigma - \overline{H}_\sigma$.*

b. *Let the segment σ, with endpoints v and w and Coxeter coordinate (p), be given and let G_σ, H_σ, $\overline{H}_\sigma$, $\overline{H}_v$ and $\overline{H}_w$ be defined as above.*
   - (i) *$G_\sigma = \sigma$, i.e., σ is the only shortest v, w-path in Λ.*
   - (ii) *For any vertex $u \in H_\sigma$, $\delta(u, v) \leq |\sigma|$ and $\delta(u, w) \leq |\sigma|$ with strict inequality in both cases whenever $u \in H_\sigma - \overline{H}_\sigma$.*

While the graph distance function δ plays an important role in our development, a second distance function will be needed. If the segment σ has Coxeter coordinates (p, q), we define the *refined length* of σ to be $p + q + \frac{|p-q|}{p+q+1}$ and denote it by $\|\sigma\|$; if σ has Coxeter coordinate (p), we define $\|\sigma\| = p + \frac{p}{p+1}$. For example, the Coxeter coordinates of a segment which has length 5 must be one of (5), $(4, 1)$, $(3, 2)$, $(2, 3)$, or $(1, 4)$ and the refined length of this segment will be $5\frac{5}{6}$, $5\frac{1}{2}$, $5\frac{1}{6}$, $5\frac{1}{6}$, or $5\frac{1}{2}$, respectively.

Let σ have Coxeter coordinates (p, q) $[(p)]$; then $\|\sigma\| = |\sigma| + \frac{|p-q|}{p+q+1}$ $(\|\sigma\| = |\sigma| + \frac{p}{p+1})$. Since $0 \leq \frac{|p-q|}{p+q+1} < 1$ $(0 < \frac{p}{p+1} < 1)$, $|\sigma| \leq \|\sigma\| < |\sigma| + 1$. It follows that $|\sigma| = \lfloor\|\sigma\|\rfloor$. And from this we can conclude that if $\|\sigma\| = \|\sigma'\|$, then $|\sigma| = |\sigma'|$. Finally, suppose that σ' has Coxeter coordinates (p', q') and that $\|\sigma\| = \|\sigma'\|$. Then $p+q = p'+q'$ and $|p-q| = |p'-q'|$. It follows that either $p' = p$ and $q' = q$ or $p' = q$ and $q' = p$. We conclude that if $\|\sigma\| = \|\sigma'\|$, then σ and σ' are congruent segments. The converse is clearly true. Leaving the special case that σ and σ' have Coxeter coordinates (p) and (p') to the reader, we have the following lemma.

LEMMA 6. *For segments σ and σ',*

i. $|\sigma| \leq \|\sigma\| < |\sigma| + 1$*;*

ii. $|\sigma| = \lfloor\|\sigma\|\rfloor$*;*

iii. *if $\|\sigma\| = \|\sigma'\|$, then $|\sigma| = |\sigma'|$;*

iv. *if $|\sigma| < |\sigma'|$, then $\|\sigma\| < \|\sigma'\|$;*

v. *σ and σ' are congruent if and only if $\|\sigma\| = \|\sigma'\|$.*

4. The signature of a plane triangulation. Let Γ be a finite plane triangulation and let P denote the set of all vertices of Γ with valence different from 6. Define the *first auxiliary graph of* Γ to be the complete graph $\mathcal{A}_1(\Gamma) = (P, K)$ and, for each $\{v, w\} \in K$, define $\omega_1(\{v, w\})$ to be the distance between v and w in Γ. Let $\mathcal{A}_2(\Gamma) = (P, E)$ be the structure graph of $\mathcal{A}_1(\Gamma), \omega_1$. $\mathcal{A}_2(\Gamma)$ is called the *second auxiliary graph of* Γ. We wish to investigate the geometry of Γ in the neighborhood of an edge of $\mathcal{A}_2(\Gamma)$. Let $\{v, w\}$ be an edge in $\mathcal{A}_2(\Gamma)$ and select a shortest path from v to w in Γ. By Lemma 1, part i, we conclude that there is no vertex $u \in P$ so that both the distance from u to v and the distance from u to w are less than $\omega_1(\{v, w\})$. In particular, all of the vertices (other than v and w) on this or any shortest path joining v and w have valence 6. Now consider this path as a subgraph of Γ. Since every vertex interior to this path has valence 6, we can trace a copy of this path in Λ such that the turns at each interior vertex are the same on the path and its copy. Label the corresponding ends of the copy by v and w. Then, by Lemma 4, this copy is a shortest v, w-path in Λ and may be identified with a segment σ in Λ as pictured in Figure 4.

Now consider the mapping from our path in Γ into Λ. We wish to extend this mapping to as large a region of Γ as is possible. Since both Λ and Γ are triangulations, we can extend this map to all vertices and edges which complete a triangle with one edge on the path or are adjacent to one vertex interior to the path. We continue to extend the domain of this map outward from the path by including adjacent vertices of degree 6. In view of Lemma 1, part i and Lemma 5, parts a(ii) and b(ii), we see that this mapping may be extended to a region of Γ which is mapped onto H_σ. We may think of the hexagonal region H_σ, pictured in Figure 4, as a region in Γ with the understanding that some of the vertices on the boundary may belong to P. We will call such a region a *hexagonal region of* Γ and we will use the same notation for this region and its various subsets as is used for their images in Λ.

Using this construction, each edge $\{v, w\}$ of $\mathcal{A}_2(\Gamma)$ may be identified with a segment in a hexagonal region of Γ. However, this identification depends on the choice of a shortest v, w-path in Γ. If σ is the segment associated with a given v, w-path, then any of the paths in G_σ will give the same hexagonal region. But Γ is a finite planar graph and perhaps there is another shortest v, w-path running "around the back." This can indeed happen, in which case we would have another segment σ' associated with the edge $\{v, w\}$ in $\mathcal{A}_2(\Gamma)$ and another hexagonal region $H_{\sigma'}$ in Γ with v and w on its boundary. When this occurs we will add another v, w-edge to $\mathcal{A}_2(\Gamma)$ associated with σ'. Thus $\mathcal{A}_2(\Gamma)$, as amended, is a multigraph and we label each edge with the Coxeter coordinates of the segments associated with that edge. Note, if σ and σ' are the segment associated with multiple edges, then $|\sigma| = |\sigma'|$.

We may actually draw this amended $\mathcal{A}_2(\Gamma)$ on the sphere by superimposing it on the given drawing of Γ: for each edge of $\mathcal{A}_2(\Gamma)$, draw in the associated segment σ as it appears in H_σ. If these segments do not cross, this will be a planar embedding of $\mathcal{A}_2(\Gamma)$. Unfortunately, some of the edges of this drawing of $\mathcal{A}_2(\Gamma)$ may cross. We solve this problem by replacing the weight function ω_1 with the weight function ω_2: for each edge e in $\mathcal{A}_2(\Gamma)$, let σ_e denote its associated segment and let $\omega_2(e) = \|\sigma_e\|$. The structure graph of $\mathcal{A}_2(\Gamma)$ with weight function ω_2 is called the *signature graph* of Γ and is denoted by $\mathcal{S}(\Gamma)$. We must keep in mind that $\mathcal{S}(\Gamma)$ could actually be a multigraph. However, we will continue to abuse notation and call it simply the signature graph of Γ. Since $\mathcal{S}(\Gamma)$ is obtained from $\mathcal{A}_2(\Gamma)$ by simply deleting some of its edges, $\mathcal{S}(\Gamma)$ inherits a natural drawing in the plane, and conveniently we have deleted enough edges to eliminate all crossings.

LEMMA 7. *For a plane triangulation* Γ*, the drawing of* $\mathcal{S}(\Gamma)$ *on the sphere described above is a planar embedding.*

Proof. Let Γ be a plane triangulation and consider the drawing of $\mathcal{S}(\Gamma)$ described above. Suppose that, in this drawing, the segments σ and σ' with endpoints $\{v, w\}$ and $\{v', w'\}$, respectively, cross. Denote the Coxeter coordinates of the segments by (p, q) and (p', q'), respectively. (We leave to the reader the similar but simpler cases where one or both of the segments have a single Coxeter coordinate.) Assume that $\|\sigma\| \geq \|\sigma'\|$; hence, $|\sigma| \geq |\sigma'|$ as well. We have drawn the hexagonal region of Γ about σ in Figure 5.

Since they belong to P, the vertices v' and w' cannot lie in the interior of H_σ. Select a shortest v', w'-path associated with the segment σ' and consider the intersection of this path with H_σ. This intersection contains a subpath which crosses σ. Let r and s denote the endpoints of this subpath, so that the clockwise order of the four points around $\overline{H}_\sigma$ is v, r, w, and s.

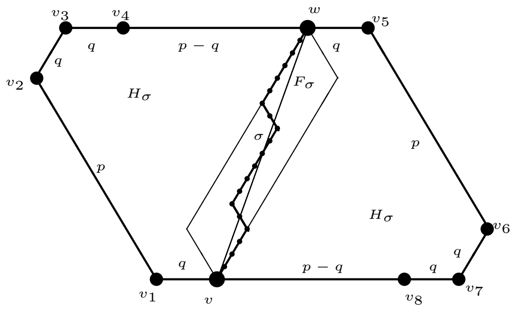


FIG. 5.

First, suppose that r lies on the section of the boundary from v_3 to w while s lies on the boundary from w to v_7. In this case, w lies on a shortest r, s-path and, hence, on a shortest v', w'-path, which is impossible. Likewise, the possibility that r lies between v and v_3 while s lies on the boundary from v to v_7 can be eliminated.

Next, suppose that r lies on the section of the boundary from v_1 to v_3 while s lies on the boundary from v_5 to v_7. In this case, $|\sigma| \geq |\sigma'| \geq \delta(r,s) \geq p+q = |\sigma|$ and equality must hold throughout. But equality can hold only if $\{v', w'\} = \{r, s\} = \{v_1, v_7\}$ or $\{v', w'\} = \{r, s\} = \{v_3, v_5\}$, and both options have already been excluded.

We conclude that one of r and s lies on the top boundary of H_σ and the other on the bottom boundary of H_σ. Thus we again have $|\sigma| \geq |\sigma'| \geq \delta(r,s) \geq p+q = |\sigma|$, and again equality must hold throughout. Thus $\{v', w'\} = \{r, s\}$ and we are free to assume $r = v'$ and $s = w'$.

Note that the segments joining v to v_4 and w to v_8 are reflections of σ and hence have the same refined length as σ. Now, if v' were to lie between v_4 and w, we easily see that the refined lengths from v' to v and v' to w are both less than $\|\sigma\|$, violating Lemma 1, part i. If v' were to lie between v_3 and v_4, the refined length from v' to v would be greater than $\|\sigma\|$, and $\|\sigma'\|$ would be even larger, violating our original assumption. All that remains is the case where v' lies between v_1 and v while w' lies between w and v_5. Again one can easily see that, in this case, $\|\sigma'\| > \|\sigma\|$, violating our original assumption. □

We may now give the formal definition of the *signature* a plane triangulation Γ. It is the signature graph $\mathcal{S}(\Gamma)$ with the planar embedding given by the natural drawing of it superimposed on Γ, with its edges labeled by the Coxeter coordinates of the associated segments and its angles labeled by the angle types given by the drawing on Γ. We illustrate this definition with an example of a geodesic dome on 62 vertices (the dual fullerene has 120 vertices).

In Figure 6, the geodesic dome is pictured in the plane and is therefore somewhat distorted. The 5-valent vertices are circled and 6 segments of the signature are drawn in as double lines (note that two of them pass through vertices of valence 6). The remaining 14 segments of the signature all have length 1 and join adjacent 5-valent vertices. The signature graph is then redrawn on the left in Figure 7 with the segment and angle labels: the two lobes are identical and, to minimize clutter, we have included the angle labels only in the top lobe and the edge labels only in the bottom lobe.

The natural question to ask is, Does the signature of a plane triangulation uniquely determine that plane triangulation? Basically we are asking if it is true that the faces of the signature can be filled in consistent with the edge and angle labels in only one way. In the next section, we show that the answer is "yes" for geodesic domes. For

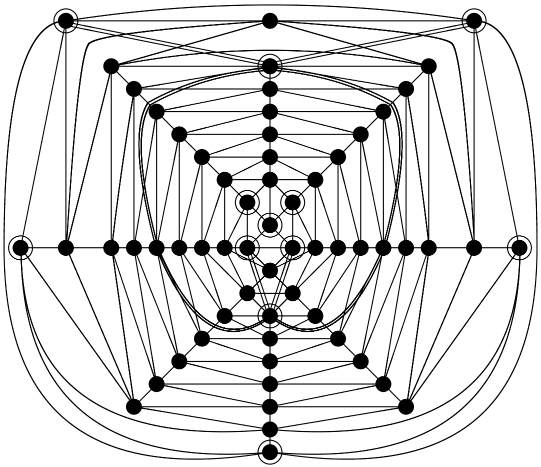


FIG. 6.

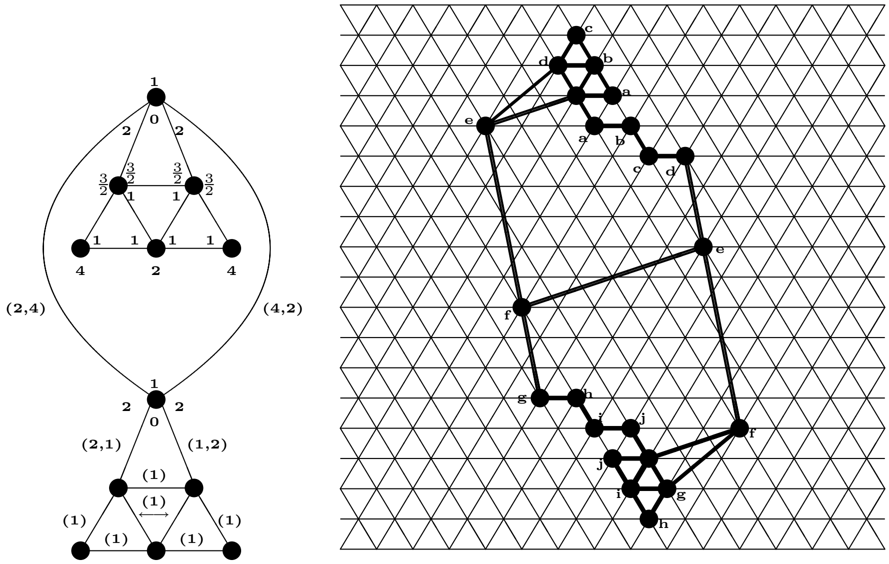


FIG. 7.

now, we simply continue with this example. On the right in Figure 7, we have drawn the faces of the signature in Λ with several of the segments identified. To build a three-dimensional model of Γ, simply cut along the unidentified segments and make the identifications indicated by the vertex labels.

5. The signature of a fullerene. The first part of the problem posed in the last section is to rebuild a plane triangulation from its signature. This boils down to filling in each face of the signature with a region from Λ that is consistent with the segment and angle labels of that face. The second part of the problem is to show that this can be done in only one way. The natural approach to filling in the faces is to select a face of the signature and then simply reconstruct its boundary in Λ

as prescribed by its segment and angle labels: denote the segments by $\sigma_1, \ldots, \sigma_k$ in clockwise order around the face; draw a segment σ'_1 in Λ congruent to σ_1; draw in a segment σ'_2 congruent to σ_2 sharing an endpoint so that the angles between σ_1 and σ_2 and σ'_1 and σ'_2 have the same type; and so on. Ultimately this is precisely what we will do; but at the outset it is not even clear that this "dead reckoning" approach will result in a closed polygonal region of Λ. To aid in our investigation we introduce some additional notation. Following Brinkmann, Friedrichs, and Nathusius [1], we define an *(m,k)-patch* to be a plane graph such that

- all faces are k-gons, except for one n-gon,
- the boundary of the n-gon is an elementary circuit and is called the boundary of the patch,
- all vertices not on the boundary have valence m while those on the boundary have valence at most m.

For a simple example of a (6,3)-patch, consider any region of Λ bounded by an elementary circuit. For a more complicated example, consider a long narrow region which curves around and overlaps itself. For the (6,3)-patch, we consider the overlapping portions of the region to be distinct. Now let Γ be a geodesic dome and consider a face of $\mathcal{S}(\Gamma)$. Replace each segment σ in the boundary of this face by a shortest path joining its endpoints that lies in G_σ. If the angle between consecutive segments σ and σ' is small, G_σ and $G_{\sigma'}$ may overlap. In this case, we select the paths in $G_\sigma \cup G_{\sigma'}$ so that they do not intersect. Next, label the vertices and edges on these paths clockwise around the face; vertices and edges which lie on paths corresponding to segments that bound the face on two sides are labeled twice, once from each side. Considering doubly labeled vertices and edges as two distinct vertices or edges, we have associated a (6,3)-patch with the given face.

By a *drawing* of a (6,3)-patch, Δ, in Λ, we mean a graph homomorphism from Δ into Λ such that distinct triangular faces sharing a common edge are mapped onto distinct triangular faces sharing a common edge. Up to an automorphism of Λ, a given (6,3)-patch has a unique drawing in Λ: select any triangular face of Δ and map into Λ; this mapping extends uniquely to its neighboring triangular faces and then to their neighbors, etc., until the entire patch is drawn. Let Δ be a (6,3)-patch with boundary Ω and let $v_0, \ldots, v_{n-1}, v_n = v_0$ be the vertices of Ω in cyclic order around the patch. The cyclic sequence of valences $\rho(v_0), \ldots, \rho(v_{n-1}, \rho(v_n) = \rho(v_0)$ is called the *boundary code* for Ω. Note that, by the definition of (m,k)-patch, $2 \leq \rho(v_i) \leq 6$, for each $i = 1, \ldots, n$. Given the boundary code of Δ, we may inductively draw this boundary in Λ: in Λ, select any two adjacent vertices v'_0 and v'_1 to be the images of v_0 and v_1; once the edge $\{v'_{i-1}, v'_i\}$ has been drawn, let v'_{i+1} be the $\rho(v_i)$th vertex adjacent to v'_i counting counterclockwise starting with v'_{i-1} and draw in $\{v'_i, v'_{i+1}\}$. At each step, the drawing of this circuit must match with the boundary of the appropriate drawing of the entire patch Δ. Thus, the drawing of Ω in Λ is unique up to an automorphism of Λ and depends only on the boundary code for Ω.

Now let Δ be a (6,3)-patch associated with a face of the signature of a geodesic dome Γ, as constructed above, and draw its boundary in Λ. The image of the boundary may then be partitioned into paths corresponding to the segments in $\mathcal{S}(\Gamma)$ from which they came. Clearly, the endpoints of each such path define a segment in Λ with the same Coxeter coordinates its corresponding segment in $\mathcal{S}(\Gamma)$. Replacing these paths by segments in Λ, results in a (perhaps overlapping) polygonal region of Λ bounded by segments corresponding to the segments bounding the given face. Furthermore, the angle labels match. We have proved the following lemma.

LEMMA 8. *Up to an automorphism of Λ, the polygonal boundary of the face of the signature of a plane triangulation has a unique drawing in Λ with the same sequence of Coxeter coordinates and angle types.*

The next question is, Is the polygonal region of a face of the signature of a geodesic dome uniquely determined by its boundary? X. Guo, P. Hansen, and M. Zheng [8] constructed two distinct $(3,6)$-patches with the same boundary sequence and his example is easily altered to produce two distinct $(6,3)$-patches with the same boundary. We say that a boundary sequence is *ambiguous* if there exist two distinct $(6,3)$-patches with the same boundary sequence. We say that the face of the signature of a plane triangulation is *ambiguous* if there exist two distinct polygonal regions with the boundary of that face. The next lemma follows at once from Lemma 8.

LEMMA 9. *A geodesic dome with a signature that admits no ambiguous faces is uniquely determined by its signature.*

If the drawing of the boundary of a (6,3)-patch is an elementary circuit, then its interior is uniquely determined and the patch is not ambiguous. Hence the drawing of an ambiguous (6,3)-patch in Λ must be self-overlapping. The drawings of faces of triangulations with vertices of valence more than 6 may well be self-overlapping and possibly ambiguous. In the remainder of this section, we sketch the proof that a face of the signature of a geodesic dome cannot yield a self-overlapping region when drawn in Λ thereby verifying:

THEOREM 1. *A geodesic dome is uniquely determined by its signature.*

The basic idea is that a face of the signature of a geodesic dome cannot curve back in itself far enough to overlap. Rather than looking at all possible faces of the signature of a geodesic dome, we consider the single face of a shortest spanning tree of the geodesic dome. This is the union of all of the faces of the signature and, if its drawing is not self-overlapping, none of the faces can have self-overlapping drawings. As a prototype self-overlapping face of a shortest spanning tree, consider the region pictured in Figure 8. There are several features to observe. First, since a spanning tree has 11 segments, this face is bounded by 22 segments. Second, in order to turn back on itself as pictured here, the face must have several angles of types greater than 3 (e.g., A, B, and C).

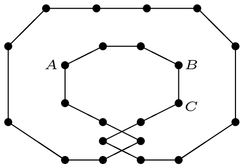


FIG. 8.

We start by eliminating the possibility of some angle types. The arguments we include here are very geometric and will be carried out in Λ. They will be valid in the geodesic dome by virtue of the fact that they will be carried out in a region of Λ corresponding to the union of overlapping hexagonal regions of signature segments in the geodesic dome.

LEMMA 10. *The signature of a geodesic dome admits no angles of types $\frac{1}{2}$ or $4\frac{1}{2}$. A shortest spanning tree of a geodesic dome admits no angles of types 0, $\frac{1}{2}$, or $4\frac{1}{2}$. Furthermore, the vertices at angles of types 4 and $3\frac{1}{2}$ in a shortest spanning tree have valence 2 while angles of type 5 occur only at pendant vertices.*

Proof. Let Γ be a geodesic dome and let segments σ with endpoints v and w and τ with endpoints u and v be segments in $\mathcal{S}(\Gamma)$. Assume that the type T of the angle between these two segments is 0 or $\frac{1}{2}$ and that $|\sigma| \geq |\tau|$. Since u cannot lie interior to the hexagonal region of σ, it must lie on the top boundary of H_σ as illustrated in Figure 9. We observe that the distance between u and w is less than the distance between v and w. Thus, both σ and τ can belong to $\mathcal{S}(\Gamma)$ only if they have the same refined length. If $T = \frac{1}{2}$, this is impossible since their Coxeter coordinates are (p, q) and $(p + q)$. Thus $\mathcal{S}(\Gamma)$ does not have an angle of type $\frac{1}{2}$. $\mathcal{S}(\Gamma)$ will include both σ and τ with $T = 0$ if τ has Coxeter coordinates (q, p). But in this case, the segment joining u and w would be considered by the shortest spanning tree algorithm before σ and τ and at most one of σ and τ could belong to a shortest spanning tree.

Since the types of the angles at a vertex in $\mathcal{S}(\Gamma)$ or a shortest spanning tree must sum to 5, an angle of type $4\frac{1}{2}$ is excluded. In a shortest spanning tree, an angle of type 4 or type $3\frac{1}{2}$ can only occur at a vertex of valence 2 across from a vertex of type 1 or type $1\frac{1}{2}$ and an angle of type 5 is excluded from a shortest spanning tree unless it is a pendant vertex. □

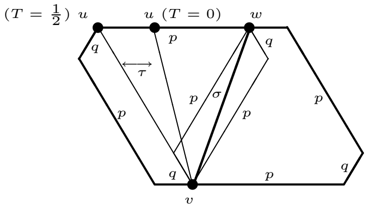


Fig. 9.

There are three basic ways in which an overlap could occur: as the result of a broad curve like that pictured in Figure 8, a broad curve in conjunction with a pendant vertex, or as the result of several pendant vertices. We consider each of these possibilities in turn.

The edges at each vertex of Λ divide the neighborhood about that vertex into six sectors. We label these sectors with the integers mod six starting with 0 at the bottom and working clockwise around the vertex; see Figure 10. Fix a shortest spanning tree and let σ_0 and σ_1 be segments of the tree with a common endpoint v_1. Suppose that, moving clockwise around the tree (counterclockwise around the face), we encounter σ_0 followed by σ_1. Suppose further that the type of the angle the segments make at v_1 is T_1 (an integer) and that σ_0 approaches v_1 through sector 0; then σ_1 leaves v_1 through sector T_1. We have illustrated this with an angle of type 4 on the left in Figure 10. If T_1 were not an integer, say $3\frac{1}{2}$, then σ_1 would leave along the edge separating sectors 3 and 4 or σ_0 could enter along the edge separating sectors 0 and 1 while σ_1 leaves through sector 4. If σ_1 leaves v_1 through sector T_1, it approaches its other endpoint through sector $T_1 - 3$. Hence, if we have a path which is a section of the boundary $\sigma_0, v_1, \ldots, v_k, \sigma_k$ with angle type T_i at v_i and if σ_0 approaches v_0 through sector 0, then σ_k approaches v_{k+1} through sector $(\sum_1^k T_i) - 3k$, (if $(\sum_1^k T_i) - 3k = T + \frac{1}{2}$, σ_k approaches v_{k+1} along the edge separating sectors T and $T + 1$). This number, $(\sum_1^k T_i) - 3k$, is called the *excess* of the path. In order to yield a self-overlapping region, this section of the boundary must make a complete clockwise change of direction and approach v_{k+1} through sector 3 or higher. That is, its excess must be at least 3: $(\sum_1^k T_i) - 3k \geq 3$ or $(\sum_1^k T_i) \geq 3k + 3$. Hence we must have several angles of type 4 or $3\frac{1}{2}$ along a portion of the tree. Since angles of types $3\frac{1}{2}$ and 4 occur only at vertices of valence 2 in our shortest spanning tree, it is not surprising that the "worst case

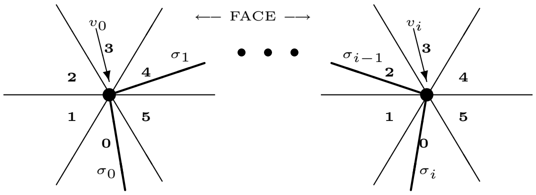


FIG. 10.

scenario" is that our shortest spanning tree is a simple path. So think of a subpath of our tree along which the angles have types 4, $3\frac{1}{2}$, and 3. We investigate this subpath by considering the angles on the other side of the path.

We are interested in constructing a path in our shortest spanning tree of length k so that the sum of the types of the angles along this path is at most $5k - (3k + 3) = 2k - 3$. In Figure 11, we have drawn an angle of type 1 at vertex v_0. Around the vertex v_{-1}, we have constructed a hexagon; any vertex in this hexagon has its distance to v_{-1} less than $|\sigma_0|$. Thus v_1 is outside (or on) the hexagon; were it inside, the segment joining v_{-1} to v_1 would have been selected in place of σ_0 or σ_1 when constructing the shortest spanning tree. If the next segment, σ_2 (from v_1 to v_2), were to make an angle of type 1, it would either (1) force v_2 to lie in the hexagon; (2) force σ_2 to be so long that it completely crosses the hexagon; (3) force σ_2 to be so short that it never intersects the hexagon. In the first case, we have the previous contradiction. In the second case, v_2 is closer to v_{-1} than to v_1, resulting in another contradiction. The third case is impossible since v_2 must lie outside the corresponding hexagon around v_1. The same argument will exclude an angle of type $1\frac{1}{2}$. Thus, the angle at v_1 is of type at least 2. At this point we note that the angle at v_0 would be of type $1\frac{1}{2}$ if σ_1 were horizontal. And the above arguments would still preclude the angle at v_1 being of type 1. However, two successive angles of type $1\frac{1}{2}$ are possible. We continue considering the case of angles of integer type and simply note that our arguments can be adapted to the cases involving angles of fractional type.

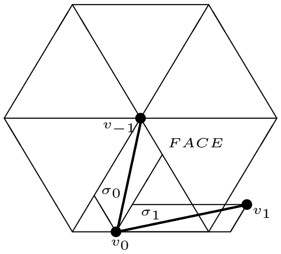


FIG. 11.

In Figure 12, we have added σ_2 with an angle of type 2 at v_1 and σ_3 with an angle of type 2 at v_2. Neither of these aid in our goal of a path with type sum $2k - 3$. It would seem that the only way to make a sharp, type 1 turn to the left is to first pull away from the hexagon and then turn toward it but not into it. To pull away, we could include a type 3 angle; but this would nullify the initial type 1 angle in the sum. The only other option is to make a segment like σ_3 much longer than the side of the hexagon. However, this won't work either. For example, the distance between v_3 and v_{-1} must be as long or longer than $|\sigma_0|$, $|\sigma_1|$, $|\sigma_2|$, and $|\sigma_3|$; otherwise the segment joining v_3 and v_{-1} would have been added to the shortest spanning tree before the longest of those segments. In short, adding a long segment at an angle of type 2 increases the size of the "forbidden hexagon" about v_{-1} for all subsequent vertices. We conclude that, as we move along a section of the boundary of the face in either direction from a vertex of type 1, we must encounter a vertex of type 3 or more before we may encounter a second vertex of type 1. A similar conclusion holds when fractional types are considered. Hence, we can never achieve a path with type sum at most $2k-3$ nor one with type sum at least $3k+3$ (excess 3 or more) using only nonpendant vertices.

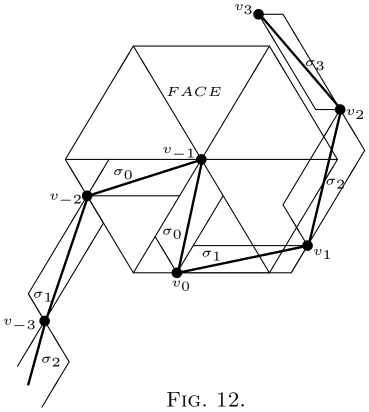


FIG. 12.

So the next question is, Can we get a little extra back curvature at a pendant vertex? The answer is "yes, but far too little for an overlap." We illustrate this by assuming that v_{-1} has valence 5. This is also pictured in Figure 12. As one can see, the face does turn back on itself—but not sharply enough to self-intersect. Because the vertices of the spanning tree have valence 5 in Γ, we have that the outside angle at v_{-1} is of type 1. If the two copies of segment σ_1 were extended to intersect, their outside angle would be of type 2 and the two copies of σ_2 make an angle of type 3. So, at each step away from v_{-1}, the two boundary paths are pulling away from one another and we conclude that no overlapping can occur near the pendant vertices.

The only remaining question is whether several pendant vertices could yield an overlap. Here there are many cases to consider. By a *branch*, we mean a single edge along with one of the components that its deletion yields. A branch with k vertices yields a section of the boundary with $2k$ edges and $2k-1$ vertices; see Figure 13 below. We can easily compute the excess of this entire section of the boundary path of the face: the sum of all angle types at these $2k-1$ vertices is $5k$; so the excess along the entire path is $5k-3(2k-1)=3-k<3$. The natural question to ask is, Could a subpath have a higher excess?

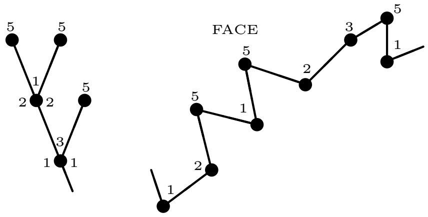


FIG. 13.

Among all possible subpaths of the face boundaries of branches, consider those with largest possible excess. Suppose that v is a pendant vertex of the branch and is interior to the subpath of the face boundary. Let $T, 5, T'$ be the angle types of the three vertices on the subpath centered at v. These three vertices contribute $T + 5 + T' - 9 = T + T' - 4$ to the excess of the path. Now delete this pendant vertex from the branch and adjust the subpath accordingly. The three vertices are replaced by a single vertex of type $T + T'$ which contributes $T + T' - 3$ to the excess. Thus the smaller branch has a subpath with a larger excess. We have carried out this reduction on the branch in Figure 13 and recorded the result in Figure 14. We conclude that the branches with a subpath having a largest possible excess have just two pendant vertices and that the subpath with largest possible excess runs from one pendant vertex to the other. Checking our examples, we see that the subpath in Figure 13 between the extreme pendant vertices has an excess of 3 while the corresponding subpath in Figure 14 has an excess of 4.

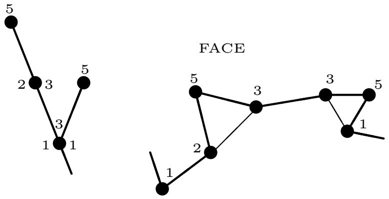


Fig. 14.

As we have noted the sum of the angle types along the entire path around a branch with k vertices is $5k$. Assume that we have a branch with just two pendant vertices. To make the excess of the subpath that runs from one pendant vertex to the other as large as is possible, we should make the sum of the types along the outside of paths from the trivalent vertex to the pendant vertices as small as possible. As we have already shown, the sum of the types along such a path will be as small as is possible when we have one angle of type 1 (or two of type $1\frac{1}{2}$) and the rest of type 2. This smallest sum is then $2(k-2)$. So the sum of the angle types along the subpath joining the two pendant vertices is $5k - 2(k-2) = 3k + 4$ and its excess is 4.

Finally, we note that a pendant vertex gives rise to two sides of an equilateral triangle along the face boundary. If we add the triangle to the face replacing the two edges by the third side of the triangle, we get a region of Λ that contains the face. We note further that making this substitution in the subpath results in decreasing the excess by 1. Carrying out this reduction throughout the entire face results in a region containing the face which has the property that no subpath of the boundary has excess 3 or more and hence is nonoverlapping. Hence the face is nonoverlapping.

6. Using the signature. It follows from Theorem 1 that the signature $\mathcal{S}(\Gamma)$ of a geodesic dome carries complete information about the geodesic dome Γ and its fullerene dual. It would be convenient if we could read some of this information directly from $\mathcal{S}(\Gamma)$. In *An Atlas of Fullerenes* [6], Fowler and Manolopoulos indicate that whether or not pentagonal faces are adjacent is an important feature of a fullerene. Clearly, the relative positions of the pentagonal faces can be read directly from its signature. Another feature of a geodesic dome/fullerene that can be easily deduced from its signature is its symmetry structure.

Let the geodesic dome Γ be given. Any automorphism α of Γ must permute the 5-valent vertices, and hence it induces a permutation of the vertices of $\mathcal{A}_1(\Gamma)$ which we also denote by α. Since α preserves distance, it is an automorphism of the weighted graph $\mathcal{A}_1(\Gamma), \omega_1$. It must then map spanning trees of $\mathcal{A}_1(\Gamma)$ onto spanning trees of $\mathcal{A}_1(\Gamma)$. It follows that α must map $\mathcal{A}_2(\Gamma)$, the structure graph of $\mathcal{A}_1(\Gamma)$, onto itself. Suppose that v and w are vertices in P joined by the segment σ in $\mathcal{A}_2(\Gamma)$. Clearly, α maps the hexagonal region of Γ determined by v and w onto the hexagonal region of Γ determined by $\alpha(v)$ and $\alpha(w)$. Thus $\alpha(\sigma)$ is congruent to σ and α preserves refined length.

Applying the above arguments to $\mathcal{A}_2(\Gamma), \omega_2$, we conclude that α induces an automorphism of its structure graph, namely $\mathcal{S}(\Gamma)$. Furthermore, if the original α is an orientation preserving automorphism of Γ, the induced α is an orientation preserving automorphism of $\mathcal{S}(\Gamma)$ that maps segments onto segments with the same Coxeter coordinates and, if the original α is an orientation reversing automorphism of Γ, the induced α is an orientation reversing automorphism of $\mathcal{S}(\Gamma)$ that maps segments onto segments with reversed Coxeter coordinates. In both cases α preserves angle types. Thus we are led to define an *automorphism* of a signature to be an automorphism of the signature graph (as a plane graph) that preserves angle types and preserves or reverses the Coxeter coordinates of the edges, according to whether it is orientation preserving or orientation reversing.

Now suppose that α is an automorphism of the signature of Γ. Since the edge and angle labels around a face determine a unique region of Λ up to an automorphism of Λ, α has a natural extension to an automorphism of Γ. Thus this mapping of the automorphism group of Γ into the automorphism group of its signature is onto. Is it one-to-one? Surprisingly, the answer is "not always." We explore these exceptional cases next.

Let α and β be automorphisms of Γ which induce the same automorphism on its signature. Since Γ is a triangulation of the plane, two automorphisms which agree on a single triangular face agree everywhere. Let v and w be any two vertices in P joined by an edge in the signature. Then α and β map the hexagonal region determined by v and w onto the hexagonal region determined by $\alpha(v) = \beta(v)$ and $\alpha(w) = \beta(w)$. If they agree on this hexagonal region, they would agree everywhere. Hence the hexagonal region must be symmetric and the two images of the v, w-hexagonal region must be reflections of one another. This is only possible if the edge is of type (p) or (p, p). Next, suppose that u, v, and w are the vertices of a path of length two in the signature. Since the image of this path must be invariant under reflection, v has valence two in $\mathcal{S}(\Gamma)$ and the angles at v are both of type $\frac{\rho}{2}$, where ρ is the valence of v in Γ. It follows that $\mathcal{S}(\Gamma)$ is either a path or a circuit, that its edge labels are of the forms (p) and (p, p) and that the angle types are equal at each vertex. If Γ is a plane triangulation with such a signature, one easily sees that both the identity and the reflection through the line or circuit of its signature induce the identity on its signature. An example of such a geodesic dome is worked out at the end of this section. We have proved the following theorem.

THEOREM 2. *Let Γ be a plane triangulation.*

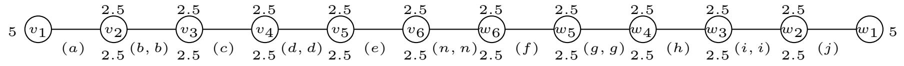
i. *If $\mathcal{S}(\Gamma)$ is not a path or circuit with the special labeling described above, then its automorphism group and the automorphism group of its signature are isomorphic.*
ii. *If $\mathcal{S}(\Gamma)$ is a path or circuit with the special labeling described above, then its automorphism group is isomorphic to the direct product of the automorphism group of its signature and the reflection through its signature.*

The next question we tackle is, How can we compute the number of vertices, edges and faces of a plane triangulation from its signature? To answer this question, consider a polygonal region with vertices $v_0, v_1, \ldots, v_n$ in counterclockwise order around the region. Fix the coordinate system with v_0 at the origin, with the horizontal edge to the right at v_0 as the unit vector in the x direction and the next edge counterclockwise as the unit vector in the y direction. The x, y-coordinates of the segment directed from the point v to the point w is simply the coordinates of w minus the coordinates of v. It should be clear that the x, y-coordinates of a segment can be computed directly from its orientation and its Coxeter coordinates. Returning to our polygonal region, the orientations of the bounding segments are determined, in order, by the orientation of the previous segment and the type of the angle between them. For $i = 1, \ldots, n$, let (x_i, y_i) denote the x, y-coordinates of the segment joining v_{i-1} and v_i. It follows that the coordinates of the vertex v_i are $(\overline{x}_i, \overline{y}_i) = (\sum_{j=1}^{i} x_j, \sum_{j=1}^{i} y_j)$. The standard formula for the area of such a polygonal region is $\frac{1}{2}\sum_{1\le i<n}(\overline{x}_i\overline{y}_{i+1} - \overline{x}_{i+1}\overline{y}_i)$. Note that a unit square in the x, y coordinate system consist of two lattice triangles and thus the area of the region is equal to the area of $\sum_{1\le i<n}(\overline{x}_i\overline{y}_{i+1} - \overline{x}_{i+1}\overline{y}_i)$ lattice triangles. Finally, substituting the x, y values for $\overline{x}, \overline{y}$ in this formula gives the area, in lattice triangles, as

$$\sum_{1\le i<j\le n} (x_i y_j - x_j y_i).$$

Using this formula, we compute the area of each face of the signature of a plane triangulation Γ. The sum t of these areas will then be the number of the triangles or faces of Γ and the numbers of edges and vertices will be given by the formulas $e = \frac{3t}{2}$ and $v = \frac{t+4}{2}$.

We close with one more example. This is the signature for one of the families of fullerenes that arose in our discussion of symmetry. The signature graph consists of a single path; the angle labels are 5 at the pendant vertices and 2.5 for all other angles; the segment labels are variables. See Figure 15.



$a + 3b + 2c + 3d + e = f + 3g + 2h + 3i + j$

$2n + d + e - a - b + f + g - i - j > max\{a, 2b, c, 2d, e, 2n, f, 2g, h, 2i, j\}$

FIG. 15.

In selecting values for the parameters the conditions listed in the figure must be satisfied. The equality ensures that the face, as laid out in Λ, actually closes. If one were to carry out this construction for a given set of values, the vertices v_1 and w_1 could be quite close to one another. In that case, (v_1, w_1) would be a segment in the signature of the fullerene we construct. So, in order to make sure that the signature with which we start is the signature of the fullerene we produce, the inequality must hold.

Using the formula for area derived above, we may compute the number of triangles in the geodesic dome or the number of vertices in the fullerene. This number turns out to be a rather complicated quadratic polynomial in these 11 variables:

$2ab + 2ac + 4ad + 2ae + 2bc + 6bd + 4be + 2cd + 2ce + 2de + 2fg + 2fh + 4fi$

$+\ 2fj + 2gh + 6gi + 4gj + 2hi + 2hj + 2ij + (d + e - a - b + f + g - i - j + 4n)s,$ where $s = a + 3b + 2c + 3d + e = f + 3g + 2h + 3i + j.$

One final observation: The fullerenes in this family are nanotubes. Select any fullerene in this family, i.e., select values for the parameters that satisfy the equality and the inequality and note that n can be increased without limit while keeping the remaining parameters fixed. In general, a nanotube will have a signature containing an edge cut set consisting of congruent segments that partition the vertices into two classes of six vertices each and have a parameter that may be enlarged independent of all other parameters. Our first example, pictured in Figures 6 and 7, is also a nanotube. The cut set consists of the double edges connecting the vertices labeled e and f in the right-hand picture of Figure 7. Replace their Coxeter coordinates with $(n, 4)$ and $(4, n)$; increasing n beyond 2 simply moves the top configuration (vertices a through e) up and to the right along the line of the tessellation.

REFERENCES

[1] G. Brinkmann, O. D. Friedrichs, and U. V. Nathusius, *Numbers of faces and boundary encodings of patches*, to appear in Graphs and Discovery, Proceedings of the DIMACS Workshop on Computer-Generated Conjectures from Graph Theoretic and Chemical Databases.

[2] D. L. D. Caspar and A. Klug, *Physical principles in the construction of regular viruses*, Cold Spring Harbor Symp. Quant. Biol., 27 (1962), pp. 1–24.

[3] H. S. M. Coxeter, *Virus macromolecules and geodesic domes*, in A Spectrum of Mathematics, J. C. Butcher, ed., Oxford University Press, Oxford, UK, 1971, pp. 98–107.

[4] P. W. Fowler, J. E. Cremona, and J. I. Steer, *Systematics of bonding in non-icosahedral carbon clusters*, Theor. Chim. Acta, 73 (1988), pp. 1–26.

[5] P. W. Fowler and J. E. Cremona, *Fullerenes containing fused triples of pentagonal rings*, J. Chem. Soc., Faraday Trans., 93 (1997), pp. 2255–2262.

[6] P. W. Fowler and D. E. Manolopoulos, *An Atlas of Fullerenes*, Clarenden Press, Oxford, UK, 1995.

[7] M. Goldberg, *A class of multi-symmetric polyhedra*, Tôhoku Math. J., 43 (1937), pp. 104–108.

[8] X. Guo, P. Hansen, and M. Zheng, *Boundary Uniqueness of Fusenes*, Tech. report G-99-37, GERAD, Montreal, Quebec, Canada, 1999.

SIAM J. Discrete Math.
Vol. 17, No. 4, pp. 615–623

INDEPENDENT SETS IN REGULAR HYPERGRAPHS AND MULTIDIMENSIONAL RUNLENGTH-LIMITED CONSTRAINTS[*]

ERIK ORDENTLICH[†] AND RON M. ROTH[‡]

Abstract. Let G be a t-uniform s-regular linear hypergraph with r vertices. It is shown that the number of independent sets $I(G)$ in G satisfies

$$\log_2 I(G) \le \frac{r}{t}\left(1 + O\left(\frac{\log^2(ts)}{s}\right)\right).$$

This leads to an improvement of a previous bound by Alon obtained for $t = 2$ (i.e., for regular ordinary graphs). It is also shown that for the Hamming graph $\mathcal{H}(n,q)$ (with vertices consisting of all n-tuples over an alphabet of size q and edges connecting pairs of vertices with Hamming distance 1),

$$\frac{\log_2 I(\mathcal{H}(n,q))}{q^n} = \frac{1}{q} + O\left(\frac{\log^2(qn)}{qn}\right).$$

The latter result is then applied to show that the Shannon capacity of the n-dimensional (d,∞)-runlength-limited (RLL) constraint converges to $1/(d+1)$ as n goes to infinity.

Key words. regular hypergraphs, Hamming graphs, multidimensional constraints, runlength-limited constraints

AMS subject classifications. 05C65, 05C69, 05A16, 68R05, 68R10, 68P30, 94A24

DOI. 10.1137/S0895480102419767

1. Introduction.

For a hypergraph G, let V_G and E_G, respectively, denote the set of vertices and set of hyperedges of G, where $E_G \subseteq \{e \subseteq V_G : |e| \ge 2\}$. For a vertex v in V_G let $N_G(v)$ denote the set of vertices that are adjacent to v in G, namely,

$$N_G(v) = \Big\{v' \in V_G \setminus \{v\} : \{v,v'\} \subseteq e \text{ for some } e \in E_G\Big\},$$

and let $\delta_G(v) = |N_G(v)|$ be the degree of v in G. An independent set in G is a subset $T \subseteq V_G$ such that $|e \cap T| \le 1$ for all $e \in E_G$. The number of independent sets in G will be denoted by $I(G)$.

A hypergraph G is *t-uniform* if each hyperedge contains t vertices, and is called *s-regular* if each vertex is contained in s hyperedges. If the intersection of any two hyperedges of G contains at most one vertex, then G is said to be *linear*. See [2].

The following theorem is the main result of this paper.

Theorem 1.1. *Let G be a t-uniform s-regular linear hypergraph with r vertices. The number of independent sets $I(G)$ in G satisfies*

$$\log_2 I(G) \le \frac{r}{t}\left(1 + O\left(\frac{\log^2(ts)}{s}\right)\right).$$

[*] Received by the editors December 16, 2002; accepted for publication (in revised form) August 17, 2003; published electronically May 17, 2004. The material in this paper was presented in part at the 2003 International Symposium on Information Theory, Yokohama, Japan, June, 2003.

http://www.siam.org/journals/sidma/17-4/41976.html

[†] Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304 (eord@hpl.hp.com).

[‡] Computer Science Department, Technion, Haifa 32000, Israel (ronny@cs.technion.ac.il). The work of this author was done while visiting Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304.

The proof of Theorem 1.1 is given in section 2, and in section 3 we present a generalization of Theorem 1.1 to uniform linear hypergraphs that are not necessarily regular.

We next present several applications of Theorem 1.1.

1.1. Regular graphs. For the special case of (undirected) regular ordinary graphs, Theorem 1.1 takes the following form.

THEOREM 1.2. *For an s-regular graph G with r vertices,*

$$\frac{\log_2 I(G)}{r} \leq \frac{1}{2} + O\left(\frac{\log^2 s}{s}\right). \tag{1}$$

Theorem 1.2 improves on the error term, $O(s^{0.1})$, which was previously obtained by Alon [1] (as shown by Kahn [7], the error term can be further improved to $O(1/s)$ when the s-regular graph G is bipartite). Unfortunately, (1) is not tight for the widely conjectured worst-case graph consisting of a disjoint union of complete bipartite graphs with degree s [1], [7]. Thus, there is still room for improvement.

1.2. Hamming graphs. Let $\mathcal{H}(n, q)$ denote the Hamming graph whose vertices are all indices $\mathbf{j} \in \{0, 1, \ldots, q-1\}^n$ and two vertices are connected by an edge if and only if they are at Hamming distance 1 apart, i.e., the vertices differ on exactly one coordinate.

The number, $I(\mathcal{H}(n, q))$, of independent sets in $\mathcal{H}(n, q)$ has received some attention in the literature ($I(\mathcal{H}(n, q))$ is also the number of codes of length n and minimum Hamming distance ≥ 2 over an alphabet of size q). The case $q = 2$ is of particular interest, and $\mathcal{H}(n, 2)$ is more commonly known as the binary Hamming hypercube. The strongest result for $q = 2$ is due to Korshunov and Sapozhenko [9] (see also [14]), who show that

$$I(\mathcal{H}(n, 2)) \sim 2\sqrt{\mathrm{e}}\, 2^{2^{n-1}},$$

where e is the base of natural logarithms; it readily follows that $2^{-n} \log_2 I(\mathcal{H}(n, 2)) = 1/2 + O(2^{-n})$.

As for general q, we have

$$\frac{\log_2 I(\mathcal{H}(n, q))}{q^n} \geq \frac{1}{q}\,, \tag{2}$$

since every subset of $\{\mathbf{j} = (j_1, j_2, \ldots, j_n) : j_1 + j_2 + \cdots + j_n \equiv 0 \pmod{q}\}$ is an independent set in $\mathcal{H}(n, q)$.

Little seems to be known about how tight the lower bound (2) is when $q > 2$. Numerical computations of $I(\mathcal{H}(n, q))$ for $q = 2, 3, 4$ and small n have been carried out [4]. We are not aware of any asymptotic analysis of $I(\mathcal{H}(n, q))$ for $q > 2$ beyond what we derive here. Specifically, we note that a subset of the Hamming graph $\mathcal{H}(n, q)$ is an independent set if and only if it is also an independent set in the q-uniform, n-regular, linear hypergraph with the same vertex set as $\mathcal{H}(n, q)$ and with hyperedges being the subsets of vertices of $\mathcal{H}(n, q)$ that agree in all but one component. Hence, by setting $r = q^n$, $s = n$, and $t = q$ in Theorem 1.1 we obtain the following result.

THEOREM 1.3. *The number of independent sets in the Hamming graph $\mathcal{H}(n, q)$ satisfies*

$$\frac{\log_2 I(\mathcal{H}(n, q))}{q^n} = \frac{1}{q} + O\left(\frac{\log^2(qn)}{qn}\right),$$

for all q.

1.3. Multidimensional runlength-limited constraints. For any n-tuple of positive integers $\mathbf{m} = (m_1, m_2, \ldots, m_n)$ let Γ be an n-dimensional $m_1 \times m_2 \times \cdots \times m_n$ binary array whose entries are indexed by n-tuples of integers

$$\mathbf{j} \in \{0, 1, \ldots, m_1{-}1\} \times \{0, 1, \ldots, m_2{-}1\} \times \cdots \times \{0, 1, \ldots, m_n{-}1\}.$$

We say that Γ satisfies the (d, ∞)-runlength-limited (RLL) constraint if and only if for any two indices $\mathbf{j}$ and $\mathbf{j}'$ that differ in only one component and differ by less than $d+1$ in that component, either $\Gamma(\mathbf{j}) = 0$ or $\Gamma(\mathbf{j}') = 0$. That is, every one-dimensional subarray of Γ satisfies the one-dimensional (d, ∞)-RLL constraint. Let $\mathcal{A}(n, d, \mathbf{m})$ be the set of all such arrays. The Shannon capacity of the n-dimensional (d, ∞)-RLL constraint is defined by

$$\mathcal{C}(n, d) = \lim_{i \to \infty} \frac{\log_2 |\mathcal{A}(n, d, \mathbf{m}^{(i)})|}{\prod_{\ell=1}^{n} m_\ell^{(i)}} \tag{3}$$

$$= \inf_{\mathbf{m}} \frac{\log_2 |\mathcal{A}(n, d, \mathbf{m})|}{\prod_{\ell=1}^{n} m_\ell}, \tag{4}$$

where $\mathbf{m}^{(i)} = (m_1^{(i)}, m_2^{(i)}, \ldots, m_n^{(i)})$ is any sequence of n-tuples of integers satisfying $\min_\ell m_\ell^{(i)} \to \infty$. That the right-hand side of (3) is independent of how the limit is taken and coincides with (4) follows from subadditivity arguments; see [6], [8].

The value $\mathcal{C}(n, d)$ equals the largest coding rate of any encoder (i.e., one-to-one mapping) from the set of finite unconstrained binary sequences into the set of (d, ∞)-RLL constrained arrays [16]. One-dimensional RLL constraints are common in magnetic and optical recording channels [10], [11], [15]. The ongoing practical interest in using multidimensional recording media (see, for example, [5] and [17]) provides the motivation for studying the values of $\mathcal{C}(n, d)$ for n greater than 1.

The following facts about $\mathcal{C}(n, d)$ are known:

1. $\mathcal{C}(1, d) = \log_2 \alpha_d$, where α_d is the positive real root of the polynomial $x^{d+1} - x^d - 1$ [15, p. 65], [16].
2. $\mathcal{C}(2, d) \sim (\log_2 d)/d$ (namely, $\lim_{d \to \infty} \mathcal{C}(2, d) \cdot (d/\log_2 d) = 1$) [8].
3. $0.5878911617 \le \mathcal{C}(2, 1) \le 0.5878911619$ [3], [13], [17].
4. $0.5225 \le \mathcal{C}(3, 1) \le 0.5269$ [13].
5. $\mathcal{C}(n, d) \ge 1/(d{+}1)$ for all n [6], [8]. This follows by further constraining the 1's in Γ to have indices $j_1, j_2, \ldots, j_n$ satisfying $j_1{+}j_2{+}\cdots{+}j_n \equiv 0 \pmod{(d{+}1)}$.

The last fact, together with the simple observation that $\mathcal{C}(n, d)$ is decreasing in n for fixed d (implied by the infimum-based specification of $\mathcal{C}(n, d)$ in (4)), raises the possibility that $\mathcal{C}(n, d)$ decreases with n all the way down to $1/(d+1)$. We next show that this is indeed the case.

Let $\mathcal{H}(n, q)$ be the Hamming graph as defined in section 1.2 and denote by $\mathbf{1}$ the n-tuple consisting of all 1's. It is not hard to see that the set of locations of 1's in any array in $\mathcal{A}(n, d, (d+1)\mathbf{1})$ corresponds to an independent set in the graph $\mathcal{H}(n, d+1)$. The reverse is also true. Hence,

$$|\mathcal{A}(n, d, (d+1)\mathbf{1})| = I(\mathcal{H}(n, d+1)).$$

On the other hand, we also have the upper bound

$$\mathcal{C}(n, d) \le \frac{\log_2 |\mathcal{A}(n, d, (d+1)\mathbf{1})|}{(d+1)^n}.$$

By Theorem 1.3 we thus get the next result.

Theorem 1.4.

$$\lim_{n\to\infty} \mathcal{C}(n,d) = \frac{1}{d+1}.$$

2. Independent sets in uniform, regular, linear hypergraphs. In this section we prove Theorem 1.1. Given a hypergraph G and a subset $Y \subseteq V_G$, let G_Y be the induced (i.e., maximally connected) subhypergraph of G on the vertices Y, that is,

$$V_{G_Y} = Y \quad \text{and} \quad E_{G_Y} = \Big\{ e \cap Y : e \in E_G, |e \cap Y| \geq 2 \Big\}.$$

Let $\mathcal{S}_i(G)$ be the set of all induced subhypergraphs of G on i vertices, namely,

$$\mathcal{S}_i(G) = \{G_Y : Y \subseteq V_G, |Y| = i\}.$$

Define $f_i(G)$ as

$$f_i(G) = \max_{H \in \mathcal{S}_i(G)} I(H). \tag{5}$$

Note that $f_1(G) = 2$, $f_{|V_G|}(G) = I(G)$, $f_i(G) \geq f_{i-1}(G)$ for $1 < i \leq |V_G|$, and

$$f_i(G) \leq 2^i. \tag{6}$$

We also define $f_0(G) = 1$ as standing for the empty independent set in an "empty" subhypergraph. Let $\mathcal{S}_i^*(G)$ denote the subset of subhypergraphs in $\mathcal{S}_i(G)$ that achieve the maximum in (5). We then have the following simple lemma.

Lemma 2.1. *Given a hypergraph G and an integer i in the range $1 \leq i \leq |V_G|$, let Δ be a nonnegative integer that satisfies $\Delta \leq \delta_H(v)$ for some vertex v of some subhypergraph $H \in \mathcal{S}_i^*(G)$. Then*

$$f_i(G) \leq f_{i-1}(G) + f_{i-\Delta-1}(G). \tag{7}$$

Proof. For any subhypergraph $H \in \mathcal{S}_i^*(G)$ and any vertex $v \in V_H$, the number of independent sets $I(H) = f_i(G)$ is equal to the sum of the number of independent sets that contain v and the number of independent sets that do not contain v. The latter is

$$I(H_{V_H \setminus \{v\}}) \leq f_{i-1}(G)$$

and the former is

$$I(H_{V_H \setminus (\{v\} \cup N_H(v))}) \leq f_{i-\delta_H(v)-1}(G).$$

The lemma follows from the fact that $f_i(G)$ is nondecreasing in i. □

The idea behind the proof of Theorem 1.1 is to start the recursion (7) with the bound $f_{i_0}(G) \leq 2^{i_0}$ for some i_0 and then proceed by bounding the result of iterating the recursion (7) up to $i = |V_G|$. The key to obtaining a good final bound is, for each i, to choose H and v to make Δ in (7) as large as possible. The extent to which this can be done depends on the structure of G.

Specializing to uniform, regular, linear hypergraphs, the following lemma provides a lower bound on the largest possible choice for Δ for each i.

LEMMA 2.2. *Let G be a t-uniform, s-regular, linear hypergraph with r vertices. Then for every $H \in \mathcal{S}_i(G)$*

$$\max_{v \in H} \delta_H(v) \geq \max\left\{\left\lceil s\Big(\frac{ti}{r} - 1\Big)\right\rceil, 0\right\}. \tag{8}$$

Proof. Fix a subhypergraph $H \in \mathcal{S}_i(G)$. We prove the lemma by counting ordered pairs of adjacent vertices in V_H in two different ways. Let

$$P = \Big\{(v, v') \in V_H \times V_H : v \neq v' \text{ and } \{v, v'\} \subseteq e \text{ for some } e \in E_G\Big\},$$

and for every $e \in E_G$ let $\beta_e = |e \cap V_H|$. Then $|P| = \sum_{e \in E_G} \beta_e(\beta_e - 1)$; that is, for each hyperedge in G we count the number of ordered pairs of elements of V_H in that hyperedge and sum this over all hyperedges. By the linearity of G each ordered pair is counted only once. Further, $\sum_{e \in E_G} \beta_e = si$ since each vertex $v \in V_H$ contributes to the sum for precisely the s hyperedges that contain it.

Since the function $(\beta_e)_{e \in E_G} \mapsto \sum_{e \in E_G} \beta_e(\beta_e - 1)$ is Schur convex [12] in the variables β_e, its minimum value subject to the constraint $\sum_{e \in E_G} \beta_e = si$ is achieved when β_e is constant-valued.[1] And, since $|E_G| = rs/t$, the minimizing β_e is $si/(rs/t) = ti/r$. Therefore,

$$\begin{aligned} |P| &\geq \min_{\beta_e} \sum_{e \in E_G} \beta_e(\beta_e - 1) \\ &= \frac{rs}{t}\frac{ti}{r}\left(\frac{ti}{r} - 1\right) \\ &= si\left(\frac{ti}{r} - 1\right). \end{aligned}$$

On the other hand, letting $\Delta = \max_{v \in H} \delta_H(v)$, we clearly have $|P| \leq \Delta |V_H| = \Delta i$. Combining the two bounds on $|P|$ and dividing by i gives (8). □

We also need the following two elementary propositions.

PROPOSITION 2.3. *The equation $x^{m+1} = x^m + 1$ has only one positive real solution α_m, which is decreasing in m. Further, $\alpha_m \leq m^{1/m}$ for $m \geq 3$.*

Proof. Write the equation as $x^m(x-1) = 1$. The left-hand side is nonpositive for x in the range $0 \leq x \leq 1$ and monotonically increasing for $x \geq 1$, implying that there is only one solution $\alpha_m > 1$. By definition $\alpha_m^m(\alpha_m - 1) = 1$ so that $\alpha_m^{m+1}(\alpha_m - 1) > 1$, implying, in turn, that $\alpha_{m+1} < \alpha_m$. Finally, for every $m \geq 3$ we have

$$x^m(x-1)|_{x = m^{1/m}} = m\left(m^{1/m} - 1\right) = m\left(\mathrm{e}^{(\log_{\mathrm{e}} m)/m} - 1\right) \geq m \cdot \frac{\log_{\mathrm{e}} m}{m} = \log_{\mathrm{e}} m > 1,$$

thus implying that $\alpha_m \leq m^{1/m}$. □

PROPOSITION 2.4. *Let $0 = m_0 < m_1 < \cdots < m_\ell$ and $0 = i_{-1} < i_0 < i_1 < \cdots < i_\ell$ be integers such that $i_{j-1} \geq m_j$ for $j = 1, 2, \ldots, \ell$, and suppose that the integer sequence $(f_i)_{i=0}^{i_\ell}$ satisfies*

$$f_i \leq f_{i-1} + f_{i-m_j-1}, \qquad 1 \leq i \leq i_\ell,$$

[1] We can obtain a tighter bound on $\max_{v \in H} \delta_H(v)$ by not ignoring the fact that β_e is integer-valued. In this case, the minimizing β_e takes on at most two values that differ by 1. The resulting bound, however, is more complicated and only slightly improves our bounds on the asymptotic number of independent sets.

where $j = j(i)$ is the unique index such that $(m_j \leq)\ i_{j-1} < i \leq i_j$. Let the real sequence $(g_i)_{i=0}^{i_\ell}$ be defined recursively by $g_0 = f_0$ and

$$g_i = \alpha_{m_j} g_{i-1}, \qquad 1 \leq i \leq i_\ell,$$

where j is such that $i_{j-1} < i \leq i_j$ and α_{m_j} is the positive real solution of $x^{m_j+1} = x^{m_j} + 1$. Then $f_i \leq g_i$ for all $0 \leq i \leq i_\ell$.

Proof. We prove by induction on i, where the induction base $i = 0$ is obvious. Turning to the induction step, suppose that $f_{i'} \leq g_{i'}$ holds for all $0 \leq i' < i$ and let j be such that $i_{j-1} < i \leq i_j$. Then

$$\begin{aligned} f_i &\leq f_{i-1} + f_{i-m_j-1} \\ &\leq g_{i-1} + g_{i-m_j-1} && (9) \\ &\leq (1 + \alpha_{m_j}^{-m_j}) g_{i-1} && (10) \\ &= \alpha_{m_j} g_{i-1} && (11) \\ &= g_i, \end{aligned}$$

where (9) follows from the induction hypothesis, (10) follows from the definition of g_i and the fact that α_{m_j} is decreasing in j (Proposition 2.3), and (11) follows from the definition of α_m. □

Proof of Theorem 1.1. Let $\Delta(i)$ equal the right-hand side of (8). For $j = 0, 1, \ldots, \ell$, let $m_0 < m_1 < \cdots < m_\ell$ be the values taken on by $\Delta(i)$ as i increases from 0 to r; clearly, $m_0 = 0$ and $m_\ell = s(t-1)$. Denote by i_j the largest i for which $\Delta(i) = m_j$. Thus

$$i_j = \left\lfloor \left(\frac{m_j}{s} + 1 \right) \frac{r}{t} \right\rfloor \tag{12}$$

and, in particular, $i_0 = \lfloor r/t \rfloor$ and $i_\ell = r$. Since $|N_H(v)| \leq i - 1$ for every vertex v in every $H \in \mathcal{S}_i(G)$ and since $|N_H(v)| \geq m_j$ for some v when $i = i_{j-1} + 1$, we have $i_{j-1} \geq m_j$. Therefore, by Lemma 2.1, the sequence $(f_i(G))_{i=0}^{i_\ell}$ with the integers m_j and i_j satisfies the assumptions of Proposition 2.4. Hence,

$$\begin{aligned} \log_2 f_r(G) &= \log_2 f_{i_\ell}(G) \\ &\leq \log_2 f_{i_0}(G) + \sum_{j=1}^{\ell} (i_j - i_{j-1}) \log_2 \alpha_{m_j} \\ &\leq i_0 + \sum_{j=1}^{\ell} (i_j - i_{j-1}) \log_2 \alpha_{m_j}, && (13) \end{aligned}$$

where α_{m_j} is the positive real solution of $x^{m_j+1} = x^{m_j} + 1$ and (13) follows from (6). Incorporating $i_j - i_{j-1} \leq (m_j - m_{j-1}) r/(ts) + 1$ (from (12)) and $i_0 \leq r/t$ into (13) yields

$$\begin{aligned} \log_2 f_r(G) &\leq \frac{r}{t} + \sum_{j=1}^{\ell} \left((m_j - m_{j-1}) \frac{r}{ts} + 1 \right) \log_2 \alpha_{m_j} \\ &\leq \frac{r}{t} + \sum_{m=1}^{m_\ell} \left(\frac{r}{ts} + 1 \right) \log_2 \alpha_m && (14) \end{aligned}$$

$$\le \frac{r}{t} + \Big(\frac{r}{ts}+1\Big)\left(2 + \sum_{m=3}^{m_\ell} \frac{\log_2 m}{m}\right) \tag{15}$$

$$\le \frac{r}{t} + \Big(\frac{r}{ts}+1\Big)\left(2 + \frac{\log_2^2(s(t-1))}{\log_2 \mathrm{e}}\right) \tag{16}$$

$$\le \frac{r}{t}\left(1 + O\Big(\frac{\log^2(ts)}{s}\Big)\right), \tag{17}$$

where (14) follows since α_m is decreasing in m, (15) follows since $\alpha_2 < \alpha_1 < 2$ and $\log_2 \alpha_m \le (1/m)\log_2 m$ for $m \ge 3$ (Proposition 2.3), and (16) follows from the fact that $\sum_{m=3}^{m_\ell} 1/m \le \log_{\mathrm{e}} m_\ell = \log_{\mathrm{e}} s(t-1)$. The bound $r \ge m_\ell = s(t-1)$ justifies (17). The proof is completed by noting that $I(G) = f_r(G)$. □

3. Nonregular hypergraphs. In this section, we generalize Theorem 1.1 to uniform linear hypergraphs that are not necessarily regular.

Given a hypergraph G, let $v_1, v_2, \ldots, v_{|V_G|}$ be a labeling of the vertices of G satisfying $\delta_G(v_1) \le \delta_G(v_2) \le \cdots \le \delta_G(v_{|V_G|})$. For $i = 1, 2, \ldots, |V_G|$ define

$$\sigma_G(i) = \frac{1}{i}\sum_{j=1}^{i} \delta_G(v_j).$$

That is, $\sigma_G(i)$ is the average degree among the i vertices with smallest degrees in G.

Following is a version of Lemma 2.2 for nonregular hypergraphs.

LEMMA 3.1. *Let G be a t-uniform linear hypergraph with r vertices. Then for all $H \in \mathcal{S}_i(G)$*

$$\max_{v \in V_H} \delta_H(v) \ge \max\left\{\left\lceil \sigma(i)\Big(\frac{ti}{r}\,\frac{\sigma(i)}{\sigma(r)} - 1\Big)\right\rceil, 0\right\}, \tag{18}$$

where $\sigma(i) = \sigma_G(i)$.

Proof. Replace $\sum_{e\in E_G} \beta_e = si$ with $\sum_{e \in E_G} \beta_e \ge i\sigma(i)$ and $|E_G| = rs/t$ with $|E_G| = r\sigma(r)/t$ in the proof of Lemma 2.2. □

For the case of s-regular hypergraphs $\sigma_G(i) = s$, so Lemma 2.2 is a special case of Lemma 3.1.

Next we combine Lemma 3.1 with Lemma 2.1 to obtain the following nonregular counterpart of Theorem 1.1.

THEOREM 3.2. *Let G be a t-uniform linear hypergraph with r vertices. The number of independent sets $I(G)$ in G satisfies*

$$\log_2 I(G) \le i_0 + \frac{r}{t}\cdot O\Big(\frac{\log^2(ts)}{s_1^2/s}\Big) \tag{19}$$

$$\le \frac{r}{t}\cdot\frac{s}{s_0}\cdot\left(1 + O\Big(\frac{\log^2(ts)}{s_1}\Big)\right) \tag{20}$$

$$\le \frac{r}{t}\cdot\frac{s}{s_0}\cdot\left(1 + O\Big(\frac{t\log^2(ts)}{s}\Big)\right), \tag{21}$$

where $s = \sigma_G(r)$ is the average degree in G, i_0 is the largest i for which $i\sigma_G(i) \le rs/t$, $s_0 = \sigma_G(i_0)$, and $s_1 = \sigma_G(i_0+1)$.

Proof. We proceed as in the proof of Theorem 1.1, but this time we let $\Delta(i)$ equal the right-hand side of (18). Also, let $0 = m_0 < m_1 < \cdots < m_\ell = s(t-1)$ be the values taken on by $\Delta(i)$ as i ranges from 0 to r.

Denote by i_j the largest i for which $\Delta(i) = m_j$; in particular, for $j = 0$ we get that i_0 is indeed the largest i for which $i\sigma_G(i) \le rs/t$, and for $j = \ell$ we get $i_\ell = r$. We note that $\sigma(i) = \sigma_G(i)$ is nondecreasing in i and hence so is $\sigma(i)(ti\sigma(i)/(rs) - 1)$ for $i \ge i_0$. Therefore, i_j is the largest integer i satisfying

$$\sigma(i)\left(\frac{ti}{r}\,\frac{\sigma(i)}{s} - 1\right) \le m_j$$

or, equivalently, the largest integer i satisfying

$$i \le \left(\frac{m_j}{\sigma(i)} + 1\right)\frac{rs}{t\sigma(i)}\ . \tag{22}$$

This characterization of i_j implies that

$$i_j > \left(\frac{m_j}{\sigma(i_j+1)} + 1\right)\frac{rs}{t\sigma(i_j+1)} - 1. \tag{23}$$

By (22) and (23) we have, for $j \ge 1$,

$$\begin{aligned} i_j - i_{j-1} &\le \frac{rs}{t}\left(\frac{m_j}{(\sigma(i_j))^2} - \frac{m_{j-1}}{(\sigma(i_{j-1}+1))^2} + \frac{1}{\sigma(i_j)} - \frac{1}{\sigma(i_{j-1}+1)}\right) + 1 \\ &\le \frac{rs}{t(\sigma(i_j))^2}\,(m_j - m_{j-1}) + 1 && (24) \\ &\le \frac{rs}{t(\sigma(i_0+1))^2}\,(m_j - m_{j-1}) + 1 && (25) \\ &= \frac{rs}{ts_1^2}\,(m_j - m_{j-1}) + 1, && (26) \end{aligned}$$

where (24) and (25) follow from the fact that $\sigma(i)$ is nondecreasing in i and that $i_0 + 1 \le i_{j-1} + 1 \le i_j$.

Inequality (13) from the proof of Theorem 1.1 applies verbatim here, and incorporating the bound (26) on $i_j - i_{j-1}$ yields

$$\begin{aligned} \log_2 f_r(G) &\le i_0 + \sum_{j=1}^{\ell}\left((m_j - m_{j-1})\frac{rs}{ts_1^2} + 1\right)\log_2 \alpha_{m_j} \\ &\le i_0 + \frac{r}{t}\cdot O\!\left(\frac{\log^2(ts)}{s_1^2/s}\right), && (27) \end{aligned}$$

where (27) follows from the same reasoning used to obtain (17): the only difference is that here $r \ge m_\ell = (t-1)s \ge (t-1)s_1^2/s$, which we need to assert that $rs/(ts_1^2)$ is bounded away from 0.

Turning to (20), by the definition of i_0 we get that $i_0 s_0 = i_0\sigma(i_0) \le rs/t$, i.e., $i_0 \le (r/t)(s/s_0)$. In addition, since $\sigma(i)$ is nondecreasing in i, we have $s_0 s_1 \le s_1^2$. Combining these two observations with (19) yields (20). Finally, the definition of i_0 also implies that $rs_1 \ge (i_0+1)s_1 > rs/t$; so, $s_1 > s/t$, which readily leads to (21). □

In general, if more is known about the behavior of $\sigma_G(i)$ for $i > i_0$, the $O(\cdot)$ term in (19) can be improved. We obtained (19) by using the pessimistic bound of $\sigma_G(i) \ge \sigma_G(i_0+1)$ for $i > i_0$. We do note, however, that (19) is tight to first order (the i_0 term) for a bipartite graph G in which the degree of any "left" vertex is smaller than the degree of any "right" vertex. In such a graph, there are necessarily more left vertices than right vertices and i_0 is easily seen to be the number of left vertices, which in turn is smaller than $\log_2 I(G)$.

REFERENCES

[1] N. Alon, *Independent sets in regular graphs and sum-free subsets of finite groups*, Israel J. Math., 73 (1991), pp. 247–256.

[2] C. Berge, *Hypergraphs: Combinatorics of Finite Sets*, North–Holland, Amsterdam, 1989.

[3] N. J. Calkin and H. S. Wilf, *The number of independent sets in a grid graph*, SIAM J. Discrete Math., 11 (1998), pp. 54–60.

[4] R. H. Hardin, Sequence numbers A027681, A027682, in The On-Line Encyclopedia of Integer Sequences, N. J. A. Sloane, ed.; available online at http://www.research.att.com/∼njas/sequences/index.html.

[5] J. F. Heanue, M. C. Bashaw, and L. Hesselink, *Volume holographic storage and retrieval of digital data*, Science, 265 (1994), pp. 749–752.

[6] H. Ito, A. Kato, A. Nagy, and K. Zeger, *Zero capacity region of multidimensional run length constraints*, Electron. J. Combin., 6 (1999), R33.

[7] J. Kahn, *An entropy approach to the hard-core model on bipartite graphs*, Combin. Probab. Comput., 10 (2001), pp. 219–237.

[8] A. Kato and K. Zeger, *On the capacity of two-dimensional run-length constrained channels*, IEEE Trans. Inform. Theory, 45 (1999), pp. 1527–1540.

[9] A. D. Korshunov and A. A. Sapozhenko, *The number of binary codes with distance* 2, Problemy Kibernet., 40 (1983), pp. 111–130 (in Russian).

[10] B. H. Marcus, R. M. Roth, and P. H. Siegel, *Constrained systems and coding for recording channels*, in Handbook of Coding Theory, V. S. Pless and W. C. Huffman, eds., Elsevier, Amsterdam, 1998, pp. 1635–1764.

[11] B. H. Marcus, P. H. Siegel, and J. K. Wolf, *Finite-state modulation codes for data storage*, IEEE J. Sel. Areas Comm., 10 (1992), pp. 5–37.

[12] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Math. Sci. Engrg. 143, Academic Press, London, 1979.

[13] Z. Nagy and K. Zeger, *Capacity bounds for the three-dimensional* $(0,1)$ *run length limited channel*, IEEE Trans. Inform. Theory, 46 (2000), pp. 1030–1033.

[14] A. A. Sapozhenko, *The number of antichains in ranked partially ordered sets*, Diskret. Mat. 1 (1989), pp. 74–93 (in Russian); translation in Discrete Math. Appl., 1 (1991), pp. 35–58.

[15] K. A. Schouhamer Immink, *Codes for Mass Data Storage Systems*, Shannon Foundation Publishers, Eindhoven, The Netherlands, 1999.

[16] C. E. Shannon, *The mathematical theory of communication*, Bell System Tech. J., 27 (1948), pp. 379–423.

[17] W. Weeks IV and R. E. Blahut, *The capacity and coding gain of certain checkerboard codes*, IEEE Trans. Inform. Theory, 44 (1998), pp. 1193–1203.

SIAM J. DISCRETE MATH.
Vol. 17, No. 4, pp. 624–633

ON DISTRIBUTIONS COMPUTABLE BY RANDOM WALKS ON GRAPHS[*]

GUY KINDLER[†] AND DAN ROMIK[‡]

Abstract. We answer a question raised by Donald E. Knuth and Andrew C. Yao, concerning the class of polynomials on $[0, 1]$ that can be realized as the distribution function of a random variable, whose binary expansion is the output of a finite state automaton driven by unbiased coin tosses. The polynomial distribution functions which can be obtained in this way are precisely those with rational coefficients, whose derivative has no irrational roots on $[0, 1]$.

We also show, strengthening a result of Knuth and Yao, that all smooth distribution functions which can be obtained by such automata are polynomials.

Key words. finite-state generator, automata, random walks on graphs, random number generation

AMS subject classifications. 65C10, 68Q05, 68Q70

DOI. 10.1137/S089548010343106X

1. Introduction.

In a 1976 paper, Knuth and Yao laid the foundations for a complexity theory of probability distribution functions [3]. They defined a computability class of distribution functions that can be "computed" by a random walk on an edge-labelled graph (this can also be thought of as a finite-state automaton driven by a sequence of random bits). They called such a graph a *finite-state generator*, or f.s.g.

Formally, an f.s.g. is a finite directed graph whose vertices are called *states*, with one designated state called the *initial state*. Some of the edges in the graph are labelled with *output strings*, which are finite binary strings. The *output* of the f.s.g. is the random sequence of bits $\alpha_1\alpha_2\alpha_3\ldots$, obtained by performing a simple random walk on its states, starting from the initial state and writing down sequentially the output strings that are encountered along the way. We identify the output with the real-valued random variable $0 \le X \le 1$ whose binary expansion is the output sequence $\alpha_1\alpha_2\alpha_3\ldots$, namely

$$X = \sum_{n=1}^{\infty} \frac{\alpha_n}{2^n}.$$

A distribution function $F(x)$ supported on $[0, 1]$ (that is, $F(0-) = 0$ and $F(1) = 1$) is called *computable* by an f.s.g., or just computable, if it can be realized as the distribution function of a random variable X generated by an f.s.g.

A natural question is to identify all computable distribution functions. Clearly there is a countable number of such functions, so the class of computable distribution functions is rather small. However, since the set of such distributions contains many Cantor-like distributions and other singular distributions which do not have a simple

[*] Received by the editors July 6, 2003; accepted for publication (in revised form) October 16, 2003; published electronically May 17, 2004.

http://www.siam.org/journals/sidma/17-4/43106.html

[†] The Einstein Institute of Mathematics, The Hebrew University of Jerusalem, Jerusalem, Israel (puzne@math.huji.ac.il).

[‡] Department of Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel (romik@wisdom.weizmann.ac.il).

description, one soon realizes that this question is (probably) too general to possess a meaningful answer.

On the other hand, if the discussion is limited to "nice" distributions, e.g., piecewise smooth distribution functions, then a beautiful algebraic connection is revealed. Knuth and Yao showed that if F is a computable distribution function, and F is real-analytic in an interval $(a, b) \subset [0, 1]$, then it must be a polynomial with rational coefficients there. (Theorem 2 below shows that it is enough to require that F be smooth in (a, b).) They constructed a family of polynomial distribution functions which are computable, but left open the question (question (v) on page 427 of [3]) of precisely which polynomials are distribution functions that can be computed by an f.s.g. The question was raised again by Yao [5], who gave some necessary conditions.

The purpose of this paper is to show that Yao's necessary conditions are sufficient. Our main result is as follows.

THEOREM 1. *A polynomial $Q(x)$ which is monotone increasing on* $[0, 1]$ *and satisfies $Q(0) = 0$, $Q(1) = 1$, can be realized as the distribution function of a random variable that is generated by an f.s.g. if and only if*

1. *$Q(x)$ has rational coefficients;*
2. *$Q'(x)$ has no irrational roots in* $[0, 1]$.

We prove two additional results. The next theorem further substantiates Knuth and Yao's claim that polynomials form the main class of interesting computable distribution functions, by showing that if a computable distribution function is smooth, then it is a polynomial. This strengthens Theorem 7.4 of [3], which shows the same for *analytic* computable distribution functions.

THEOREM 2. *Let F be a computable distribution function. If F is infinitely differentiable on an interval* $(a, b) \subset [0, 1]$, *then F is a polynomial there.*

The last theorem investigates some structural properties of f.s.g.'s that compute nonsmooth distributions. Recall that any distribution function F can be decomposed into a mixture

$$F = \lambda F_{\text{ac}} + (1 - \lambda) F_{\text{sing}}, \qquad 0 \le \lambda \le 1, \tag{1}$$

of an absolutely continuous distribution function F_{ac} and a singular distribution function F_{sing} (for the purpose of this paper we include the atomic part of F in F_{sing}; see also the comment in section 5). λ is determined uniquely, and if $0 < \lambda < 1$, namely if F is not purely singular or absolutely continuous, then F_{ac} and F_{sing} are also determined uniquely (otherwise, one of them is trivially not).

THEOREM 3. *Let $F(x)$ be a computable distribution function, let $F = \lambda F_{\text{ac}} + (1 - \lambda) F_{\text{sing}}$ be the decomposition of F as in* (1), *and assume that $0 < \lambda < 1$. Then λ is rational, and F_{ac} and F_{sing} are both computable.*

In the proof of Theorem 3 it is shown that, essentially, the contributions to the absolutely continuous and singular parts, respectively, come from different parts of the f.s.g. which do not interact.

Remarks. The above definition of an f.s.g. is a slight variation on those of [3, 5] but is easily seen to be equivalent, in the sense that the class of computable distribution functions is the same. In [3, 5] it was required that the outdegree of each vertex in the graph be 2 (this restriction is natural when an f.s.g. is interpreted as a coin-tossing automaton). In section 3 below, we use another equivalent variation on the f.s.g. model.

Our paper was inspired by the recent work of Mossel and Peres [4], which deals with questions somewhat similar to ours. Mossel and Peres characterize the class of functions $f : (0,1) \to (0,1)$ for which there exists a finite-state automaton whose input is a sequence of random bits with bias p and whose output is a single random bit with bias $f(p)$. Those functions are precisely the rational functions of p with rational coefficients.

Structure of the paper. In the next section we prove Theorem 1. The "only if" part was already proved in [3] and [5]. For the "if" part, we rely essentially on Knuth and Yao's construction involving the order statistics of uniform random variables. It is amusing that order statistics should play a distinguished role in this problem, and that in fact by taking scalings and rational mixtures of polynomials constructed using order statistics, one obtains the most general class of constructible polynomials.

In section 3 we prove Theorem 3. In section 4 we prove Theorem 2. In section 5 we give an example of a computable distribution function which is absolutely continuous but whose density is everywhere locally unbounded, and discuss related open problems.

2. Proof of Theorem 1. It will be convenient, in the proof of Theorem 1, to deal with density functions rather than cumulative distribution functions. Let $\mathcal{D}$ be the set of piecewise polynomial density functions on $[0,1]$. Let $\mathcal{C}$ be those elements $q(x) \in \mathcal{D}$ such that the corresponding cumulative distribution function $Q(x) = \int_0^x q(t)dt$ is computable. The elements of $\mathcal{C}$ are called computable (piecewise polynomial) densities.

The following theorem summarizes Knuth and Yao's constructions of computable densities.

Theorem 4 (see [3]). (i) *If $0 \le a < b \le 1$ are rational, then the uniform density on $[a,b]$ is computable.*

(ii) *If $0 \le a < b \le 1$ are rational, then the density*

$$f(x) = \frac{(n+1)!}{k!(n-k)!(b-a)^{n+1}} \quad (x-a)^k (b-x)^{n-k} \mathbf{1}_{[a,b]}(x)$$

of the $(k+1)$th order statistic of $n+1$ independent random variables distributed uniformly on $[a,b]$ is computable.

(iii) *If $f_1, f_2, \ldots, f_n$ are computable densities, then any rational mixture of the form $f = \sum_{i=1}^n a_i f_i$, where $0 < a_i \in \mathbb{Q}$, $\sum_i a_i = 1$, is also computable.*

Let $q \in \mathcal{D}$ be a polynomial density function such that $Q(x) = \int_0^x q(t)dt$ satisfies the conditions of Theorem 1. In terms of q, this simply means that q has rational coefficients, and no irrational roots in $[0,1]$. Our aim is to show that q is computable. Let $0 = r_0 < r_1 < r_2 < \cdots < r_{k-1} < r_k = 1$ be the roots of q in $[0,1]$, together with 0 and 1 if they are not roots. In view of Theorem 4(iii), it is enough to show that each of the densities

$$q_i(x) = \frac{1}{\int_{r_i}^{r_{i+1}} q(t)dt} \; q(x)\mathbf{1}_{[r_i,r_{i+1}]}(x), \qquad i = 0,1,2,\ldots,k-1$$

(the density q conditioned on the interval $[r_i, r_{i+1}]$), is computable. This is because q is then a mixture of the q_i with rational coefficients.

Now fix i, $0 \le i \le k-1$. q_i is a density that is 0 outside the interval $[r_i, r_{i+1}]$. Inside this interval, q_i has the form

$$q_i(x) = c(x-r_i)^j (r_{i+1}-x)^l h(x), \tag{2}$$

where $c \in \mathbb{Q} \cap (0,\infty)$, $j, l \ge 0$, and $h(x)$ is a polynomial with rational coefficients that is *strictly positive* on $[r_i, r_{i+1}]$, and integrates to 1 there. Our claim now relies on the following result.

Proposition 1. *$h(x)$ can be expressed as a rational mixture (a convex combination with rational coefficients) of polynomials which have the form*

$$c(x-t_1)^{v_1}(x-t_2)^{v_2}\cdots(x-t_{m-1})^{v_{m-1}}(-x+t_m)^{v_m} \tag{3}$$

for some rational $r_i \le t_1 < t_2 < \cdots < t_m \le r_{i+1}$ *and which integrate to* 1 *on* $[r_i, r_{i+1}]$*—the constant c takes care of this and is therefore necessarily rational. The powers $v_1, v_2, \ldots, v_m$ above must be even, with the exception that if $t_1 = r_i$, then v_1 can be odd, and if $t_m = r_{i+1}$, v_m can be odd (this is why the last term in* (3) *is written differently from the other terms).*

Proposition 1 implies our claim that q_i is computable. To see this, let f be a polynomial density on $[r_i, r_{i+1}]$ which has the form (3) (note that not only h, but also q_i is a mixture of such polynomials, by (2)). We prove that f is computable by showing that its restriction to each subinterval $[t_j, t_{j+1}]$ (normalized to have integral 1) is a computable density. On $[t_j, t_{j+1}]$, write f as

$$f(x) = c\,[(x-t_j)+(t_j-t_1)]^{v_1}[(x-t_j)+(t_j-t_2)]^{v_2}\cdots(x-t_j)^{v_j}$$

$$\cdot(t_{j+1}-x)^{v_{j+1}}[(t_{j+1}-x)+(t_{j+2}-t_{j+1})]^{v_{j+2}}\cdots[(t_{j+1}-x)+(t_m-t_{j+1})]^{v_m}.$$

Now expand out the products, observing that $t_j - t_1$, $t_j - t_2, \ldots,$ $t_j - t_{j-1}$, $t_{j+2} - t_{j+1}, \ldots,$ $t_m - t_{j+1}$ are all positive rational numbers. This gives a representation of f as a rational mixture of polynomials proportional to $(x-t_j)^\alpha(t_{j+1}-x)^\beta$; hence by Theorem 4(ii), (iii), the restriction of f to $[t_j, t_{j+1}]$ is computable.

Our goal is now to prove Proposition 1. We start by discussing how a nonnegative polynomial density on an interval can be represented as a convex combination of polynomial densities which are *not necessarily rational.*

Lemma 1. *Let $C_n[a,b]$ be the closed convex set of nonnegative polynomials of degree at most n on an interval $[a,b]$ that integrate to* 1 *there. Then $C_n[a,b]$ is a compact set, and its extreme points are precisely the polynomials in $C_n[a,b]$ of degree exactly n which have the form* (3) *for some $a \le t_1 < t_2 < \cdots < t_m \le b$ and positive even $v_1, v_2, \ldots, v_m$ (again, with the exception that if $t_1 = a$, v_1 can be odd, and if $t_m = b$, v_m can be odd).*

As was indicated to us by a referee, a proof of Lemma 1 appears in the 1953 paper by Karlin and Shapley [1, Theorem 9.2, p. 28]. We include the proof here for completeness.

Proof. Recall that a bounded closed set within a finite-dimensional normed space is compact. The space of n-degree polynomials is finite-dimensional, and it can be equipped with the norm defined by $||f|| = \int_a^b |f(x)|dx$. The set $C_n[a,b]$ is bounded with respect to this norm (all of its elements have norm 1), and it is obviously closed; hence it is compact.

Now let $f \in C_n[a,b]$ be a polynomial of degree n with all n roots (counting multiplicities) in the interval $[a,b]$ (the evenness of the multiplicities is automatic from the nonnegativity requirement), and suppose $f = \alpha g + (1-\alpha)h$, where $g, h \in C_n[a,b]$ and $0 < \alpha < 1$. From positivity we have that wherever f vanishes, g and h must also vanish with at least the same order, so they share the same n roots as f and are therefore equal to it, since they are both of degree at most n and integrate to 1. Thus f is an extreme point of $C_n[a,b]$.

Conversely, if $f \in C_n[a,b]$ does *not* have n roots in the interval $[a,b]$, then it can be represented as

$$f(x) = c(x-t_1)^{v_1}(x-t_2)^{v_2}\cdots(-x+t_m)^{v_m}\cdot g(x) =: w(x)\cdot g(x),$$

where $a \le t_1 < t_2 < \cdots < t_m \le b$, the sum of the multiplicities $\deg w = \sum_i v_i$ is strictly less than n, the constant $c > 0$ is chosen so that $g \in C_n[a,b]$, and g has no roots in $[a,b]$. Now, either of two cases must hold: if g is a constant, then $\deg f = \deg w < n$, and then

$$f(x) = \left(\int_a^b \frac{t-a}{b-a} f(t)dt\right)\left(\left(\int_a^b \frac{t-a}{b-a} f(t)dt\right)^{-1} \frac{x-a}{b-a} f(x)\right)$$

$$+ \left(\int_a^b \frac{b-t}{b-a} f(t)dt\right)\left(\left(\int_a^b \frac{b-t}{b-a} f(t)dt\right)^{-1} \frac{b-x}{b-a} f(x)\right)$$

represents f as a convex combination of two unequal polynomials in $C_n[a,b]$. Otherwise, $\deg g \ge 1$, in which case, letting $\epsilon = \min_{x\in[a,b]} g(x)$, the equation

$$f(x) = \left(\frac{\int_a^b w(t)(g(t)-\epsilon)dt}{2}\right) \cdot \frac{w(x)(g(x)-\epsilon)}{\int_a^b w(t)(g(t)-\epsilon)dt}$$

$$+ \left(\frac{\int_a^b w(t)(g(t)+\epsilon)dt}{2}\right) \cdot \frac{w(x)(g(x)+\epsilon)}{\int_a^b w(t)(g(t)+\epsilon)dt}$$

represents f as a convex combination of two polynomials in $C_n[a,b]$ which (because $\deg g \ge 1$) are not equal. Therefore f is not an extreme point of $C_n[a,b]$. □

Proof of Proposition 1. First, note that it is enough to show that $h(x)$ can be expressed as a mixture of polynomials of the form (3), without insisting on a *rational* mixture: this is since for a linear system of equations with rational coefficients, the set of rational solutions is dense in the set of real solutions.

Now, the idea of the proof is to first use Lemma 1 to represent $h(x)$ as a convex combination of polynomials of the form (3), with $r_i \le t_1 < t_2 < \cdots < t_m \le r_{i+1}$ not necessarily rational. The t_i's are then slightly perturbed to make them rational.

Proposition 1 follows from the three lemmas below as follows. First, note that since $h(x)$ has no roots, it is actually an interior point of $C_n[r_i, r_{i+1}]$, where $n = \deg h$ (we consider $C_n[r_i, r_{i+1}]$ as a subset of the affine vector space of polynomials of degree at most n that integrate to 1 on $[r_i, r_{i+1}]$). By Lemma 2, this implies that $h(x)$ is also in the interior of the convex hull of some finite set P of polynomials of the form (3). According to Lemma 3 the polynomials in P may be perturbed slightly while maintaining $h(x)$ in the interior of their convex hull. Finally, Lemma 4 implies that these perturbations can be chosen so that the roots of the polynomials become rational.

LEMMA 2. *For a set B, denote by B° the interior of B. Let K be a compact convex body in a finite-dimensional vector space V, and let $n = \dim(V)$. Then for every interior point $x \in K^\circ$ there exist extreme points $y_1, \dots, y_m$ of K such that $x \in \mathrm{Conv}^\circ(y_1, \dots, y_m)$. The number of points, m, is at most $2n$.*

LEMMA 3. *Let $x, y_1, \dots, y_n$ be points in a finite-dimensional vector space V. Suppose that $x \in \mathrm{Conv}^\circ(y_1, \dots, y_n)$. Then there exists a neighborhood $\mathcal{U}$ of $0 \in V$ with the following property. If $z_1, \dots, z_n \in V$ satisfy $z_i - y_i \in \mathcal{U}$ for all i, then $x \in \mathrm{Conv}^\circ(z_1, \dots, z_n)$.*

LEMMA 4. *The set of extreme points in $C_n[r_i, r_{i+1}]$, all of whose roots are rational, is dense in the set of extreme points of $C_n[r_i, r_{i+1}]$ (with the obvious topology).*

Lemmas 3 and 4 are obvious; hence we prove only Lemma 2. Note that the bound $2n$ on the number of required extreme points in Lemma 2 is tight, as can be seen by taking $x = 0$ and $K = \mathrm{Conv}(\pm e_1, \dots, \pm e_n)$.

Proof of Lemma 2. Assume that $K^\circ \neq \emptyset$, so that there will be something to prove. Without loss of generality, assume that $x = 0$. We choose a basis $y_1, \dots, y_n$ for V whose elements are extreme points of K, as follows.

Take y_1 to be any extreme point of K ($y_1 \neq 0$). Having chosen $y_1, \dots, y_i$ for $i < n$, we set $H_i = \mathrm{span}(y_1, \dots, y_i)$. Since K contains a neighborhood of 0, it cannot be contained in H_i. Therefore there exists an extreme point y_{i+1} of K, satisfying $y_{i+1} \notin H_i$ (for example, there exists an extreme point maximizing the convex function $\mathrm{dist}(\cdot, H_i)$, where dist is computed according to some norm on V. Recall that a convex function defined on a closed convex body always attains its maximum on some extreme point). This process obviously yields a basis for V.

Take z to be the intersection of the boundary of K with the ray $\{t \cdot (-y_1 - y_2 - \cdots - y_n) : t > 0\}$. Obviously, the convex hull of $y_1, \dots, y_n, z$ contains a neighborhood of 0. Now let H_z be an affine hyperplane supporting K at z. The intersection of K with H_z is a convex body in a vector space of dimension $\leq n - 1$, and therefore by Carathéodory's theorem (see [2]) z is a convex combination of at most n extreme points $y_{n+1}, \dots, y_m$ in it. Since these are also extreme points of K, and since, obviously, $\mathrm{Conv}(y_1, \dots, y_n, z) \subseteq \mathrm{Conv}(y_1, \dots, y_m)$, the proof is complete. □

3. Proof of Theorem 3.

In the next two sections, we slightly modify our model of f.s.g.'s to an equivalent model. In the modified model, the outgoing edges are labelled with transition probabilities, which are arbitrary rational numbers in $(0, 1]$ (and which sum to 1 for any given state). The random walk which is performed is then a weighted random walk with these transition probabilities. We also require every edge to be labelled with a single output bit.

The equivalence of the two models is simple, and was noted in [3, pp. 421–422].

Let S be the set of states of such a modified f.s.g. An alternative description of the f.s.g. is in terms of the matrix of transition probabilities, which we denote by

$$A = (p_{s \to s'})_{s, s' \in S}.$$

A is a Markov transition matrix with rational entries, and is decomposed as the sum of two substochastic matrices with rational entries

$$A = A_0 + A_1,$$

where A_0 has nonzero entries for those edges whose output label is "0" and A_1 has nonzero entries for those edges with output label "1." Specifying the f.s.g. is equivalent to specifying the matrices A_0, A_1 and the initial state s_0.

Let $F = \lambda F_{\mathrm{ac}} + (1 - \lambda) F_{\mathrm{sing}}$ be as in Theorem 3, and suppose that S is the set of states of a given f.s.g. that computes F, with initial state $s_0 \in S$. For any state $s \in S$, let F^s be the distribution function generated by the same f.s.g. with the initial state replaced by s. Thus, $F = F^{s_0}$. Thinking of the F^s as measures on $[0, 1]$, we denote for any Borel subset $B \subset [0, 1]$

$$F(B) = \int_B dF(x).$$

A state $s \in S$ is said to be of absolutely continuous (a.c.) type if F^s is an absolutely continuous measure. Call s of singular type (or just singular) if F^s is a singular measure. Call s *pure* if it is either absolutely continuous or singular.

LEMMA 5. 1. *If $s \in S$ is pure, and $s' \in S$ is a state such that there exists a path in the graph of the f.s.g. leading from s to s', then s' is pure and of the same type as s.*

2. *If the graph of the f.s.g. is strongly connected (namely, there is a path from any state to any other state), then all the states are pure (and are therefore of the same type by part* 1*).*

Proof. Let $\mu = (F^s)_{s \in S}$ be the vector-valued measure whose coordinates are the measures F^s. The definition of the f.s.g. and the measures F^s can be translated into the following system of equations satisfied by μ: for any Borel subset $B \subset [0,1]$ and any state $s \in S$,

$$F^s(B) = \sum_{s \xrightarrow{0} s'} p_{s \to s'} F^{s'}(2B \cap [0,1]) + \sum_{s \xrightarrow{1} s'} p_{s \to s'} F^{s'}((2B-1) \cap [0,1]),$$

with $s \xrightarrow{\alpha} s'$ meaning that s has an outgoing edge to s', labelled by the output bit α. In matrix notation, this can be written as

$$\mu(B) = A_0 \mu(2B \cap [0,1]) + A_1 \mu((2B-1) \cap [0,1]), \tag{4}$$

where μ is thought of as a column vector.

Now let s be an a.c. state, and let s' be a state such that $s \xrightarrow{\alpha} s'$, with α being either 0 or 1. Then for any Borel set $B \subset [0,1]$ which has Lebesgue measure 0, we have

$$0 = F^s((B+\alpha)/2) \geq p_{s \to s'} F^{s'}(B). \tag{5}$$

Therefore $F^{s'}$ is also a.c. Similarly, if s is singular, then, taking $C \subset [0,1]$ a set of Lebesgue measure 0 such that $F^s(C) = 1$, and $B = [0,1] \setminus (2C - \alpha)$, again (5) holds. This proves that s' is singular.

For part 2 of the lemma, observe first that (4) uniquely determines a vector $\mu = (F^s)_{s \in S}$ of probability measures on $[0,1]$—this is equivalent to saying that the output of the f.s.g. is a well-defined random variable. Now, for any state $s \in S$, let $F^s = \lambda(s) F^s_{\text{ac}} + (1 - \lambda(s)) F^s_{\text{sing}}$ be the decomposition of F^s into a mixture of an a.c. probability measure and a singular probability measure. We claim that, when the graph of the f.s.g. is strongly connected, the coefficients $\lambda(s)$ in these decompositions are all equal. This is because, by (4), $\lambda(s)$ is a harmonic function on this (finite) graph and is therefore constant (take as the subset B in (4) the union of the supports of all the measures F^s_{sing}).

So if $0 < \lambda = \lambda(s) < 1$, then we have shown that

$$\mu = \lambda \mu_{\text{ac}} + (1 - \lambda) \mu_{\text{sing}},$$

where μ_{ac} and μ_{sing} are vector-valued measures, each coordinate of which is a probability measure. But then, both μ_{ac} and μ_{sing} are easily seen to be solutions of (4), and therefore we have found two different (in fact, mutually singular) solutions to (4), in contradiction to the fact that (4) has exactly one solution. Therefore λ must be 0 or 1, and all the states are pure. □

COROLLARY. *$\lambda = \lambda(s_0)$ is rational, and $F^{s_0}_{\text{ac}}$, $F^{s_0}_{\text{sing}}$ are computable.*

Proof. The states of the f.s.g. decompose into strongly connected components. Call a strongly connected component *terminal* if it has no edges going out to other strongly connected components. Clearly, with probability one the random walk on the states must end up in a terminal component. Looking at a terminal component as a sub-f.s.g., Lemma 5 implies that its states must be pure, since the measures F_s for the sub-f.s.g. are the same as for the original one. Call a strongly connected component with pure states either a.c. or singular, according to the type of its states.

The above discussion leads to an identification of the mixture coefficient $\lambda(s_0)$: it is simply the probability that the random walk eventually ends up in one of the a.c. terminal components. This probability is clearly rational, as it can be represented as the solution of a (well-posed) system of linear equations with rational coefficients. From the discussion it is also easy to see how to build an f.s.g. that computes F_{ac}: simply delete any edges going into singular components, and renormalize the transition probabilities so that the sum of the probabilities of outgoing edges for any state is 1. (In other words, the new f.s.g. is the old f.s.g. conditioned never to go into a singular component.) A similar construction replacing the words "singular" and "a.c." computes F_{sing}. □

4. Proof of Theorem 2. Let F be a distribution function, computable by a given f.s.g. with state set S and initial state s_0, which is infinitely differentiable on an interval $(a,b) \subset [0,1]$. Let $x \in (a,b)$ be a dyadic number, i.e., of the form $x = k/2^m$ for some integers $m \geq 1$, $0 \leq k < 2^m$. For every $n \geq m$, we shall apply (4) n times repeatedly, starting with the set

$$B = \left[x, x + \frac{1}{2^n}\right].$$

Some notation will help: if the binary expansion of x is $x = 0.\alpha_1\alpha_2\ldots\alpha_n$ (the last $n-m$ digits are 0), and for $\alpha \in \{0,1\}$ we denote by T_α the set operation

$$T_\alpha(C) = 2C - \alpha, \qquad C \subset [0,1],$$

then applying (4) successively gives the vector equation string

$$\begin{aligned}\mu(B) &= A_{\alpha_1}\mu(T_{\alpha_1}(B)) = A_{\alpha_1}A_{\alpha_2}\mu(T_{\alpha_2}\circ T_{\alpha_1}(B)) = \cdots \\ &= A_{\alpha_1}A_{\alpha_2}\ldots A_{\alpha_{n-1}}A_{\alpha_n}\mu(T_{\alpha_n}\circ\cdots\circ T_{\alpha_1}(B)) \\ &= (A_{\alpha_1}A_{\alpha_2}\ldots A_{\alpha_{m-1}}A_{\alpha_m})(A_{\alpha_{m+1}}\ldots A_{\alpha_n})\mu([0,1]) \\ &= (A_{\alpha_1}A_{\alpha_2}\ldots A_{\alpha_{m-1}}A_{\alpha_m})A_0^{n-m}\mu([0,1]) =: A_x A_0^{n-m}\mu([0,1]) = A_x A_0^{n-m}\mathbf{1}.\end{aligned}$$

Here, $\mathbf{1}$ is the vector of all ones $(1)_{s\in S}$, and A_x is, as above, the matrix with rational entries obtained by multiplying A_0's and A_1's corresponding to the m bits in the binary expansion of x. Taking the s_0th coordinate in the above equation, we obtain

$$F(B) = F\left(x + \frac{1}{2^n}\right) - F(x) = \mathbf{1}_{s_0}^\top A_x A_0^{n-m}\mathbf{1}, \tag{6}$$

where $\mathbf{1}_{s_0}$ is the state vector all of whose coordinates are 0 except the s_0th coordinate, which is 1. Now observe that, since F is infinitely differentiable at x, then for any j the left-hand side of (6) has the asymptotic expansion as $n \to \infty$

$$F\left(x + \frac{1}{2^n}\right) - F(x) = F'(x)\cdot\frac{1}{2^n} + \frac{F''(x)}{2}\cdot\frac{1}{2^{2n}} + \cdots + \frac{F^{(j)}(x)}{j!}\cdot\frac{1}{2^{jn}} + O\left(\frac{1}{2^{(j+1)n}}\right).$$

For the right-hand side, on the other hand, we can write down a complete expansion in terms of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_l$ of the matrix A_0: clearly it must be of the form

$$\sum_{i=1}^{l} c_i \lambda_i^n p_{\lambda_i}(n)$$

for some constants c_i and polynomials $p_{\lambda_i}(t)$ derived from x, the matrices A_x, A_0, and the vectors $\mathbf{1}_{s_0}, \mathbf{1}$ (the polynomials p_{λ_i} appear when A_0 is not diagonalizable).

Equating the two expansions as $n \to \infty$, we conclude the following.

LEMMA 6. *At any dyadic $x \in (a, b)$, F can have at most $|S|$ nonzero derivatives.*

The proof of Theorem 2 will be complete once we prove the following simple lemma.

LEMMA 7. *Let F be an infinitely differentiable function on an interval (a, b), and let $\mathcal{D} \subseteq (a, b)$ be a dense subset, such that in every point $x \in \mathcal{D}$, F has at most l nonzero derivatives. Then F is a polynomial on (a, b) of degree at most l.*

Proof. Suppose for the sake of contradiction that F is not a polynomial of degree at most l. Then there exists a point $x \in (a, b)$, where its $(l + 1)$th derivative is nonzero. By continuity, there exists a subsegment $(a_{l+1}, b_{l+1}) \subseteq (a, b)$ where the $(l + 1)$th derivative of F is nonzero.

The lth derivative is strictly monotone on (a_{l+1}, b_{l+1}), and hence it crosses zero at most once. Hence there is a subsegment $(a_l, b_l) \subseteq (a_{l+1}, b_{l+1})$, where both the lth derivative and the $(l + 1)$th derivative are nonzero. Continuing by induction, one obtains an interval $(a_1, b_1) \subseteq (a, b)$, where all derivatives up to order $(l + 1)$ are nonzero. This is a contradiction to the assumption that F has at most l nonzero derivatives in every point of $\mathcal{D}$ (since $\mathcal{D} \cap (a_1, b_1) \neq \emptyset$). □

5. Open problems. Several natural questions arise from the paper:

1. Our proof of Theorem 1, which is presented in a somewhat abstract form, can easily be translated into an algorithm for constructing an f.s.g. that computes a given polynomial distribution function F. The resulting algorithm, however, seems to generate extremely large f.s.g.'s, as a function of the degree of the given polynomial and the denominators of its coefficients.
   It is interesting to determine the complexity class of finding the smallest f.s.g. that computes a given polynomial. Another interesting question is to give a sharp bound on the number of states required to compute a polynomial of given parameters.
2. One may consider the same questions that are discussed here, in the case of pushdown automata. Partial results in this direction are given in [5].
3. It may be of interest to investigate the computable distribution functions among the absolutely continuous (and not necessarily smooth) distributions. This class contains some peculiar specimens, such as the distribution computed by the f.s.g. in Figure 1 below. This distribution is absolutely continuous, yet its density function is nowhere locally bounded.
4. A sufficient condition for the distribution function F computed by a given f.s.g. to be a.c. is that any terminal component of the graph (considered as a sub-f.s.g.) outputs a uniform distribution on $[0, 1]$ starting from any of its states. Is this condition necessary?
5. Characterize all the *atomic* computable distributions.

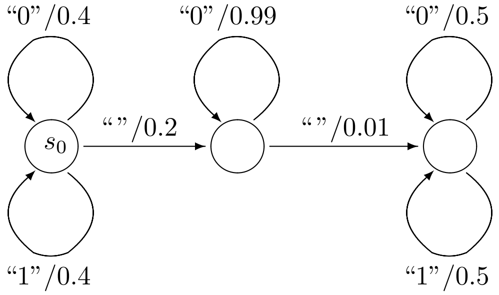


Fig. 1. *An f.s.g. generating a nowhere bounded density.*

REFERENCES

[1] S. Karlin and L. S. Shapley, *Geometry of Moment Spaces*, Mem. Amer. Math. Soc. 12, AMS, Providence, RI, 1953.

[2] P. J. Kelly and M. L. Weiss, *Geometry and Convexity*, Wiley, New York, 1979.

[3] D. E. Knuth and A. C. Yao, *The complexity of nonuniform random number generation*, in Algorithms and Complexity: New Directions and Recent Results, J. F. Traub, ed., Academic Press, New York, 1976, pp. 375–428.

[4] E. Mossel and Y. Peres, *New coins from old: Computing with unknown bias*, to appear.

[5] A. C. Yao, *Context-free grammars and random number generation*, in Combinatorial Algorithms on Words, NATO Adv. Sci. Inst. Ser. F. Comput. Systems Sci. 12, A. Apostolico and Z. Galil, eds., Springer-Verlag, Berlin, 1985, pp. 357–361.

SIAM J. Discrete Math.
Vol. 17, No. 4, pp. 634–660

LISTEN TO YOUR NEIGHBORS: HOW (NOT) TO REACH A CONSENSUS[*]

NABIL H. MUSTAFA[†] AND ALEKSANDAR PEKEČ[‡]

Abstract. We study the following rather generic communication/coordination/computation problem: In a finite network of agents, each initially having one of the two possible states, can the majority initial state be computed and agreed upon by means of local computation only? We study an iterative synchronous application of the local majority rule and describe the architecture of networks that are always capable of reaching the consensus on the majority initial state of its agents. In particular, we show that, for any truly local network of agents, there are instances in which the network is not capable of reaching such a consensus. Thus, every truly local computational approach that requires reaching a consensus is not failure-free.

Key words. consensus, graph processes, local majority rule

AMS subject classifications. 68R10, 05C75, 68M10, 68M12

DOI. 10.1137/S0895480102408213

1. Introduction.

Attempting to solve a complex problem by a simultaneous coordinated activity of local agents is an idea that arises naturally in a variety of contexts. For example, this idea is fundamental in frameworks as diverse as distributed computing and neural networks. While methods of local computation and decision-making are often effective in dealing with complex tasks, the successful implementation of such methods often raises a new breed of problems related to coordination and communication of local agents.

In this paper we study the following rather generic communication/coordination/computation problem: In a finite network of agents, each initially having one of the two possible states, can the majority initial state be computed and agreed upon by means of local computation only? Our simple model assumes bidirectional communication between agents (agent i knows agent j's state if and only if agent j knows agent i's state) and a synchronous, discrete time, democratic local decision-making procedure (an agent changes its state at time $t + 1$ if and only if the majority of agents it communicates with are in the opposite state at time t). We describe the architecture of networks that are always capable of reaching the consensus on the majority initial state of its agents. In particular, we show that, for any truly local network of agents, there are instances in which the network is not capable of reaching such a consensus. Thus, every local computational approach that requires reaching consensus among agents' results is not failure-free.

A precise formulation of the model will be given in the next section. Informally, the vertices of a graph $G = (V, E)$ represent the agents, and the edges of G represent all (bidirectional) communication links between pairs of agents. Initially, at time $t = 0$, each agent is in one of the two possible states, e.g., colored red or blue (voted Yes or No, having value 0 or 1, ...). Then the local majority rule is applied synchronously and iteratively as follows: An agent has different colors at time t and $t+1$ if and only

[*] Received by the editors May 23, 2002; accepted for publication (in revised form) October 20, 2003; published electronically May 17, 2004. Part of the work presented here has appeared in an extended abstract form in [31].
http://www.siam.org/journals/sidma/17-4/40821.html

[†] Department of Computer Science, Duke University, Durham, NC 27708 (nabil@cs.duke.edu).
[‡] The Fuqua School of Business, Duke University, Durham, NC 27708 (pekec@duke.edu).

if the agent's color at time t is not a majority color in the agent's neighborhood in G at time t. We call this discrete time, memoryless, synchronous dynamic process the *local majority process* on G.

The local majority process (and some of its natural extensions) has been studied in frameworks as diverse as social influence [19, 11, 5, 38, 39, 40] and neural networks [18, 17, 15, 16]. Recently, the local majority process has reappeared (under the name *polling process*) in several papers motivated by certain distributed computing problems [36, 2, 10, 9, 20, 21, 22, 27, 32, 33]. In fact, Peleg [35] points out several areas of distributed computing in which our model could be relevant.[1] These are areas that revolve around the idea of eliminating the damage caused by failed processors, or at least restricting their influence, by maintaining replicated copies of crucial data and performing a simple voting procedure among the participating processors whenever faults occur, with the goal of adopting the values stored at the majority of the processors as the correct data. Relevant work can be found in classical problems of agreement and consensus [1, 26, 3, 8], system-level diagnosis [42, 34, 6], distributed database management [4, 23], quorum systems [14, 12, 41, 37, 43], and fault-local mending [25, 24].

To see a concrete example, suppose that all processors in a distributed network collectively store some value and suppose that this value is distorted in some of the processors (distortions could be due to various reasons, even due to a fundamental imprecise nature of floating point operations). The goal is to restore the correct value in all of the processors by means of local communication only, in particular, by triggering the local majority process. For example, if stored distortions are due to a rounding error (rounding up or down), a desirable feature would be for all processors to accept the rounded value which is stored in the majority of processors. Which network structures allow for successful restoration of the (global) majority value in all of the processors?

A natural question to ask is, When does the local majority process ensure that all agents reach a consensus on the initial majority state? We will say that G is a *majority consensus computer* (m.c.c.) if, for any set of initial states (there are 2^n such sets), the local majority process simultaneously brings all agents into the state that was the initial majority state. Note that, according to the local majority process, once all agents are in the same state, no agent will change its state ever after. All of the recent papers dealing with the local majority process and its modifications [36, 2, 10, 9, 20, 21, 22, 27, 32, 33] investigated how poorly the local majority process (and its variations) could miscalculate the initial majority (on a specific class of graphs).[2] In contrast to these results, we are interested in graphs which are immune to miscalculations in the local majority process—the focus of this paper is on m.c.c.'s and the investigation of their structure.

Since being an m.c.c. is seemingly a very strong property, one would expect that a sort of an impossibility theorem holds. As will be shown, the situation is not that simple, and the full characterization of m.c.c.'s remains an open problem. However, our results demonstrate in several ways that the nonlocality is an inherent property

[1] We believe that the potential applicability of the local majority process goes beyond classical distributed computing problems. For example, anyone interested in data aggregation by means of local computation/communication only should be interested in this model (at least as a starting point towards possible more complex models).

[2] For example, Berger [2] has shown that for every n there exists a G on at least n vertices and the set of states such that only 18 vertices are in one state and the rest are in the other, yet the local majority process forces all vertices to simultaneously end up in the initial minority state.

of every m.c.c. Thus, reaching a consensus on the majority is a truly nonlocal task in the sense that a natural local computation procedure is failure-free only if computing local majority is essentially as complex as computing global majority.

As already mentioned, the local majority process is precisely formulated in the next section. Furthermore, we review some known properties of the model and formally define the class of graphs that we call m.c.c.'s. We end section 2 by stating and proving several basic properties of m.c.c.'s.

In section 3 we explore the structure of m.c.c.'s. For example, in this section we show that every such m.c.c. must have a trivial min-cut, a nonunique max-cut, and a diameter of at most four, and show that for any vertex v in an m.c.c. the set of vertices that are neighbors of v or neighbors of the neighbors of v is a majority-making set (i.e., has more than half of the vertices of G).

In section 4 we study highly connected graphs, i.e., those with the minimum degree of $n-3$, and show that if such a graph is an m.c.c., then there must exist a "truly global" vertex, which we call a *master* (that is, a vertex connected to every other vertex in the graph). Furthermore, we give full characterization of m.c.c.'s with $\delta(G) \geq n-3$ and present an algorithm to decide whether or not a given such graph is an m.c.c. Also, we show that there exist m.c.c.'s on n vertices, where n is odd, with exactly k masters for every positive k except for $k = (n-3)/2$.

Some generalizations of our model and relaxations of the definition of m.c.c. are presented and discussed in section 5. This section includes emulation results showing how our model can be used to study seemingly more complex models.

In section 6 we discuss some assumptions of our model and try to illustrate why our model is a natural one to study.

We close the paper with a brief summary of our results and directions for further research.

2. Democratic consensus computers. A standard graph theoretic notation is used throughout the paper. Cardinality of the set S is denoted by $|S|$ and the complement of the set is denoted by S^c. $G = (V, E)$ denotes an undirected, simple, finite graph G with the vertex set V, $|V| = n$, and the edge set E (i.e., $E \subseteq \{S \subseteq V : |S| = 2\}$). We say that the vertices u and v are adjacent or *neighbors* in G if and only if $\{u, v\} \in E$. The *neighborhood* of a vertex v in the set $S \subseteq V$ is the set of the neighbors of v that are in S, $N_S(v) := \{s \in S : \{v, s\} \in E\}$. Note that $N_S(v) = N_V(v) \cap S$. The *degree* of a vertex v in S, denoted $deg_S(v)$, is the number of neighbors of v that are in S, i.e., $deg_S(v) = |N_S(v)|$. In the rest of the text, we will omit the subscript when $S = V$; i.e., we will refer to $N_V(v)$ as $N(v)$ and to $deg_V(v)$ as $deg(v)$. The maximum degree of a vertex in G is denoted by Δ, where $\Delta = \Delta(G) = \max\{deg(v) : v \in V\}$, and the minimum degree is denoted by δ, where $\delta = \delta(G) = \min\{deg(v) : v \in V\}$.

Given a pair of nonempty $S, T \subseteq V$, let $E(S, T) = \{\{s, t\} \in E : s \in S,\ t \in T\}$. Recall that $G = (V, E)$ is bipartite with bipartition $S \cup S^c = V$ if $E(S, S^c) = E$. A pair of nonempty sets S, S^c defines an (S, S^c)-*cut* represented by $E(S, S^c)$. A cut is *trivial* if either S or S^c is a one element set. An (S, S^c)-cut is a *min-cut* if $|E(S, S^c)| \leq |E(T, T^c)|$ for all pairs of nonempty sets $T, T^c \subset V$. Similarly, an (S, S^c)-cut is a *max-cut* if $|E(S, S^c)| \geq |E(T, T^c)|$ for all pairs of nonempty sets $T, T^c \subset V$.

A graph $H = (V', E')$ is a *subgraph* of G, denoted $H \subseteq G$, if $V' \subseteq V$ and $E' \subseteq E$. We also use the notation $G \setminus H = (V, E \setminus E')$. The following graphs on n vertices are denoted in the standard way: the complete graph K_n, the path P_n, and the cycle C_n. A $K_k \subseteq G$ is a *clique* (or a k-*clique*) in G. The *complement* of G is denoted by $G^c = (V, E^c) = K_n \setminus G$. G is *connected* if, for every pair of vertices $u, v \in V$, there exists a path $P \subseteq G$ containing both u and v. Otherwise, G is *disconnected*. A *connected component* H of G is a maximal connected subgraph $H \subseteq G$. The *distance* between two vertices u and v in G, denoted $dist(u, v)$, is the smallest k for which there exists $P_{k+1} \subseteq G$ containing both u and v (this might not be defined in a disconnected graph). The *diameter* of a connected graph G is $diam(G) = \max\{dist(u, v) : u, v \in V\}$.

Some nonstandard terminology follows: A vertex v is a *master* if $deg(v) = n - 1$ (i.e., v is adjacent to every other vertex). We also say that v is a *k-master* if $deg(v) = n - 1 - k$ (i.e., v is adjacent to all but k other vertices). Note that 0-master and master are equivalent notions, and we will use them interchangeably throughout the rest of the text.

In our model, all agents and communication links in the system are represented by a graph G in a natural way. That is, the vertices of G are in a one-to-one correspondence with the agents, and the edges of G correspond to an adjacency relation among the agents.

A *coloring* of the graph G, $c^t : V \to \{0, 1\}$ defines an assignment of binary values (colors) to the vertices of G at time t. We use the notation $c_v^t := c^t(v)$ to denote the color of a vertex v at time t. The notation $sum(c^t) := \sum_{v \in V} c_v^t$ will also be useful. A color assigned to more than $|V|/2$ vertices at a time t is called the *majority color* of the coloring c^t and denoted by $maj(c^t)$. Thus, $maj(c^t) = 1$ if and only if $sum(c^t) > n/2$, and $maj(c^t) = 0$ if and only if $sum(c^t) < n/2$. Note that $maj(c^t)$ is not defined if $|V|$ is even and c^t defines an equipartition of V, i.e., if $sum(c^t) = n/2$. A coloring c^t is a *consensus* if it is constant, i.e., if all the vertices of G have the same binary values (colors). Thus, c^t is a consensus if and only if $c_v^t = maj(c^t)$ for all $v \in V$. We will sometimes abuse the notation and write $c^t = 0$ or $c^t = 1$ for consensus in color 0 and 1, respectively. Another abuse of notation is $(1 - c^t)$ denoting the coloring obtained from c^t by changing the color of every vertex; i.e., for every $v \in V$ and coloring c^t, $(1 - c^t)(v) = 1 - c^t(v)$.

Note that in our model, $c^t(v)$ corresponds to the state of an agent, represented by v, at time t.

The main object of our study is the *local majority process* $LMP(G, c^0)$, a discrete time process on G that is based on the iterative application of the local majority rule. The process is completely defined by G and the *initial coloring* c^0. For every $t = 0, 1, 2, \ldots$, the coloring c^{t+1} is derived by applying the *local majority rule* on $N(v)$ for each vertex in G:

$$c_v^{t+1} = \begin{cases} c_v^t & \text{if } |\{w \in N(v) : c_w^t = c_v^t\}| \geq |N(v)|/2, \\ 1 - c_v^t & \text{if } |\{w \in N(v) : c_w^t \neq c_v^t\}| > |N(v)|/2. \end{cases} \tag{2.1}$$

The local majority rule simply states that, at the next discrete time step, the color assigned to a vertex v will be the color of the majority of its neighbors. Note that an even degree vertex will retain its color whenever exactly half (or more) of its neighbors have the same color. The above rule also implies that the local majority rule is executed simultaneously for all the vertices. The change from c^t to c^{t+1} is called a *global update* of G at time $t + 1$, while the change of the color of a particular vertex v from c_v^t to c_v^{t+1} is called a *local update*. We say that there is a *majority switch* at time $t + 1$ if $maj(c^t) \neq maj(c^{t+1})$.

Note that if c^t is a consensus, then $c^{t+k} = c^t$ for all positive integers k. If, for some positive integer t, c^t is a consensus, then we say that G *reaches consensus* for

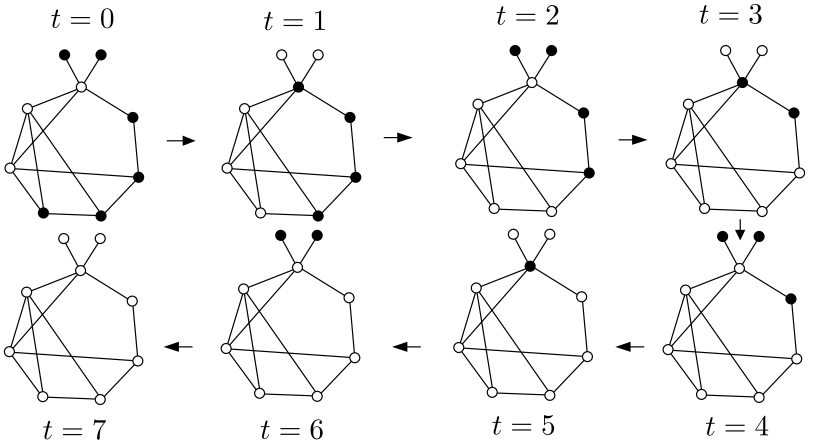


FIG. 2.1. *An example of a local majority process that reaches consensus, but not in the initial majority.*

c^0. If G reaches consensus c^t for coloring c^0 and $c^t = maj(c^0)$, then we say that the $LMP(G, c^0)$ *correctly computes the initial majority* and that G admits a *majority consensus* for the initial coloring c^0. See Figure 2.1 for an example of $LMP(G, c^0)$ in which a consensus is reached but the initial majority is computed incorrectly.

A graph G is an m.c.c. if, for every coloring c^0, $LMP(G, c^0)$ correctly computes the initial majority. In other words, G is an m.c.c. if G admits majority consensus for all of the 2^n possible initial colorings. Note that for every graph with an even number of vertices there exists a c^0 where $maj(c^0)$ is not defined. Thus, G can be an m.c.c. only if it has an odd number of vertices. Therefore, throughout the rest of the paper we **assume that n is odd**.

Our first observation about m.c.c.'s is the following proposition.

PROPOSITION 2.1. *Let G be an m.c.c. and let c^0 be an initial coloring of G. Then there are no majority switches for $LMP(G, c^0)$; i.e., $maj(c^t) = maj(c^0)$ for $t = 0, 1, 2, \ldots$.*

Proof. The local majority process reaches consensus on $maj(c^0)$ and on $maj(d^0)$ for the initial colorings c^0 and $d^0 = c^t$, respectively, because G is an m.c.c. Since $d^{t'} = c^{t+t'}$, $maj(c^t) = maj(d^0) = maj(c^0)$. □

The related research in the area of neural networks and models of social influence was geared towards finding properties of the local majority process rather than towards finding specific graphs (i.e., network architectures) having certain desirable properties. In particular, we use results about the behavior of the sequence $c^0, c^1, c^2, \ldots$. There are only 2^n possible colorings, and c^{t+1} is a function of G and c^t; thus the sequence $c^0, c^1, c^2, \ldots$ must become periodic; i.e., there exist positive integers t_0 and k such that $c^{t+k} = c^t$ for every $t \geq t_0$. Obviously, the period k and t_0 are not larger than 2^n. Somewhat surprisingly, the period can be only one or two and the minimal such t_0 is always smaller than $|E|$. We first state the original result from the neural network literature[3] and then show how this result applies to our model.

[3] Theorem 2.2 is not the most general result. Many variations can be found in a rather comprehensive collection of results related to dynamic behavior of neural and automata networks by Goles and Martinez [16]. The *period is either one or two* property holds in models beyond the symmetric neural network model. For example, dynamical systems with more general threshold functions and

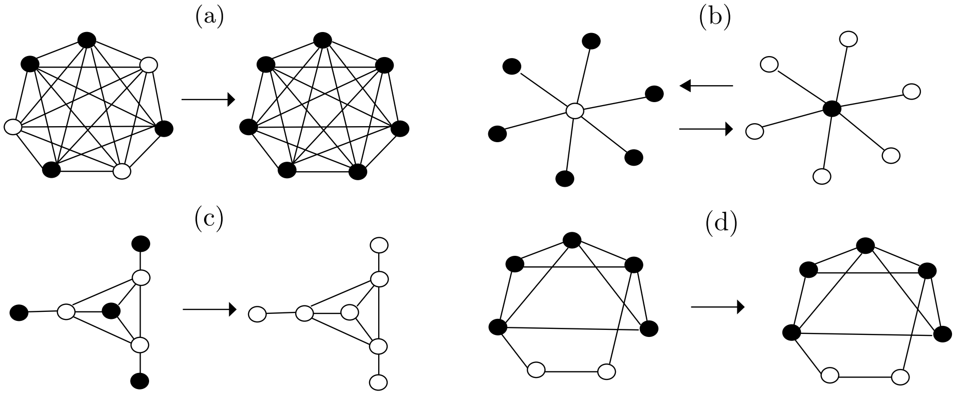


FIG. 2.2. *Examples of the local majority process.* (a) *Complete graph reaches majority consensus after one iteration,* (b) *infinite process alternating between two colorings,* (c) *reaching consensus in the minority value, and* (d) *process ending in a nonconsensus coloring.*

THEOREM 2.2 (Goles and Olivos [18]; Goles [15]). *Let $A = [a_{ij}]$ be an $n \times n$ matrix and $b \in \mathbf{R}^n$. For any $c^0 \in \{0,1\}^n$ define a dynamic process by $c_i^{t+1} = p[Ac^t - b]_i$, where $p(x) = 1$ if $x \geq 0$ and $p(x) = 0$ if $x < 0$.*

If A is symmetric, there exists t_0 such that $c^t = c^{t+2}$ for all $t \geq t_0$. Furthermore, if A is integer valued and $b = \frac{1}{2}A(1,\ldots,1)^T$ has no integral coordinates, then t_0 can be chosen so that $t_0 \leq |(\sum_{i,j}|a_{ij}|) - n|/2$.

COROLLARY 2.3. *Consider the sequence $c^0, c^1, c^2, \ldots$ defined by the local majority process on G with initial coloring c^0, $LMP(G, c^0)$. Then there exists $t_0 \leq |E|$ such that $c^t = c^{t+2}$ for every $t \geq t_0$.*

Proof. Let A be a slightly modified adjacency matrix of G; i.e., let $A = [a_{uv}]$ be defined with

$$a_{uv} = \begin{cases} 1 & \text{if } \{u,v\} \in E, \\ 1 & \text{if } u = v \text{ and } deg(v) \text{ is even}, \\ 0 & \text{otherwise.} \end{cases}$$

It is straightforward to check that, for any c^0, a dynamic process from Theorem 2.2 with A as defined is exactly $LMP(G, c^0)$. Note that A is a zero-one symmetric matrix: $a_{uv} = a_{vu}$ since $\{u,v\} = \{v,u\}$. Further note that $\sum_{u,v}|a_{uv}| = 2|E| + |\{v : deg(v) \text{ is even}\}|$. Thus, t^0 from Theorem 2.2 can be chosen so that $t_0 \leq (2|E| + |\{v : deg(v) \text{ is even}\}| - n)/2 \leq |E|$. □

Many of our results will be based on the *"period is at most two"* property. See Figure 2.2 for examples of various outcomes of a one-step application of the local majority process on graphs. Note that consensus on any value might not be possible and the process could be infinite with either period one or two, as examples (b) and (d) illustrate.

Next we show that a monotonicity property with respect to the structure of the coloring holds in the local majority process. As the next lemma shows, if at time t the color of some set of vertices is changed from $1 - i$ to i and colors of all other vertices remain the same, then, at any later time $t' \geq t$ the number of vertices of color i is at least as large as it would be without the change that was executed at time t.

allowing for more than two possible colors are studied in [38, 39, 40], while sufficient conditions for the property in the case of LMP on infinite graphs were studied in [29, 28, 30].

LEMMA 2.4. *Let $V_i(c^t) = \{v \in V : c_v^t = i\}$, $i = 0, 1$, where c^t is a coloring of $G = (V, E)$. If there exists $i \in \{0, 1\}$ and colorings c^t and $d^{t'}$ such that $V_i(c^t) \subseteq V_i(d^{t'})$, then $V_i(c^{t+k}) \subseteq V_i(d^{t'+k})$ for $k = 0, 1, 2, \ldots$.*

Proof. The proof is by induction on k. If $k = 0$, there is nothing to prove. Suppose $V_i(c^{t+k}) \subseteq V_i(d^{t'+k})$. We have to show that, for every $v \in V$, $c_v^{t+k+1} = i \Rightarrow d_v^{t'+k+1} = i$. It follows from the assumption that $N(v) \cap V_i(c^{t+k}) \subseteq N(v) \cap V_i(d^{t'+k})$ and, in particular, $c_v^{t+k} = i \Rightarrow d_v^{t'+k} = i$. Hence, if $c_v^{t+k+1} = i$ because $|N(v) \cap V_i(c^{t+k})| > |N(v)|/2$, then $|N(v) \cap V_i(d^{t'+k})| > |N(v)|/2$ also, and $d_v^{t'+k+1} = i$. If $c_v^{t+k+1} = i$ because $c_v^{t+k} = i$ and $|N(v) \cap V_i(c^{t+k})| = |N(v)|/2$, then $d_v^{t'+k} = i$ and $|N(v) \cap V_i(d^{t'+k})| \geq |N(v)|/2$, which shows that $d_v^{t'+k+1} = i$. □

According to the definition, in order to check whether G is an m.c.c., one would have to check whether G admits majority consensus for all 2^n possible initial colorings c^0. However, because of the monotonicity property described in Lemma 2.4, it suffices to consider only colorings c^0 such that $sum(c^0) = (n+1)/2$ (there are $\binom{n}{(n+1)/2} = O(2^n/\sqrt{n})$ such colorings).

THEOREM 2.5. *Suppose G admits majority consensus for any coloring c^0 such that $sum(c^0) = (n+1)/2$. Then G is an m.c.c.*

Proof. By symmetry, if G admits majority consensus for all c^0 with $maj(c^0) = 1$, then G admits majority consensus for all c^0 with $maj(c^0) = 0$ also (because $maj(1 - c^0) = 1 - maj(c^0) = 1$).

Let d^0 be a coloring with $maj(d^0) = 1$, i.e., $sum(d^0) \geq (n+1)/2$. Thus, there exists a coloring c^0 with $sum(c^0) = (n+1)/2$ such that c^0 and d^0 satisfy conditions of Lemma 2.4 with $i = 1$ (e.g., construct c^0 from d^0 by changing the color of any $sum(d^0) - sum(c^0)$ vertices w such that $d_w^0 = 1$). Since G admits majority consensus for c^0, using the terminology of Lemma 2.4, there exists t such that $V_1(c^t) = V$ and, by the lemma, $V_1(c^t) \subseteq V_1(d^t)$. Therefore, d^t is a consensus with $maj(d^t) = 1$, which shows that G admits majority consensus for d^0. □

Remark. Unfortunately, it is not true that adding an edge to or deleting an edge from an m.c.c. G preserves the property "majority concensus computer." In other words, if G is an m.c.c., $G + e$ might not be. Similarly, if G is not an m.c.c., $G - e$ could be. For example, consider

$$(K_n)^c \subset K_n \setminus P_{n-1} \subset K_n \setminus P_{(n+1)/2} \subset K_n,$$

where n is odd. We later show that $(K_n)^c$ is not an m.c.c. (Corollary 3.3), $K_n \setminus P_{n-1}$ is an m.c.c. (Theorem 4.13), $K_n \setminus P_{(n+1)/2}$ is not an m.c.c. ((b) of Proposition 3.1), and that K_n is an m.c.c. ((a) of Proposition 3.1). Thus, the graph property *"majority concensus computer"* is *not* monotone in the sense that addition or deletion of an edge in G does not preserve the property.

We close this section by showing that masters in G compute majority instantly, i.e., the color of a master at time $t+1$ is $maj(c^t)$. In general, the larger the difference between the majority and minority color of c^t, the smaller the degree of v needed to ensure $c_v^{t+1} = maj(c^t)$. Recall that a vertex v is a k-master if $deg(v) = n - (k+1)$.

PROPOSITION 2.6. *If v is a master in G, then $c_v^{t+1} = maj(c^t)$. More generally, if v is a k-master in G and $|sum(c^t) - n/2| \geq (k+1)/2$, then $c_v^{t+1} = maj(c^t)$.*

Proof. Note that $|\{w \in V : c_w^t = 1 - maj(c^t)\}| \leq deg(v)/2$ implies $c^{t+1}(v) = maj(c^t)$ (at time t, at most half of v's neighbors have color $1 - maj(c^t)$; if each of the two colors is the color of exactly half of v's neighbors, then no other vertex has color $1 - maj(c^t)$ and, in particular, $c_v^t = maj(c^t)$, and the local majority process ensures

$c_v^{t+1} = c_v^t = maj(c^t)$). In order to complete the proof, note that $|sum(c^t) - n/2| \geq (k+1)/2$ is equivalent to $|\{w \in V : c_w^t = 1 - maj(c^t)\}| \leq (n - (k+1))/2$. □

3. Structural properties. Let's start by presenting a class of graphs that are m.c.c.'s and a class of graphs that are not m.c.c.'s.

Proposition 3.1.

(a) *A graph G with more than $n/2$ masters is an m.c.c.*

(b) *A graph G with exactly $(n-1)/2$ masters is not an m.c.c.*

Proof. First suppose that G has more than $n/2$ masters. Then, by Proposition 2.6, for any c^0 and any master $v \in V$, $c_v^1 = maj(c^0)$. It follows that $maj(c^1) = maj(c^0)$. Thus, $c_v^2 = maj(c^1) = maj(c^0)$ (the first equality follows from Proposition 2.6 with $t = 1$). Also, $c_w^2 = maj(c^0)$ because masters are the majority of w's neighbors and, as already observed, $c_v^1 = maj(c^0)$ for every master v. Hence, c^2 is the consensus in color $maj(c^0)$, and (a) follows.

In order to prove (b), let G be a graph with exactly $(n-1)/2$ masters and let $c_v^0 = 0$ if v is master and $c_w^0 = 1$ if v is not a master. Note that $maj(c^0) = 1$. Every $w \in V$ that is not a master is connected to all $(n-1)/2$ masters (all having color 0 at time $t = 0$) and is connected to at most $(n+1)/2 - 2 = (n-3)/2$ vertices that are not masters (w is not connected to itself and to at least one more vertex u because w is not a master; u is not a master either because it is not connected to w). Thus, $c_w^1 = 0$ and $maj(c^1) = 0 \neq maj(c^0)$. Hence, by Proposition 2.1, G is not an m.c.c. □

Next we give a characterization of m.c.c.'s that indicates a way towards a static representation in the form of existence of a particular partition of the vertices of G.

Theorem 3.2. *G is* not *an m.c.c. if and only if at least one of the following holds:*

(a) *There exists a coloring c^0 such that $maj(c^0) \neq maj(c^1)$.*

(b) *There exists a partition of V into four sets A_0, A_1, B_0, B_1 satisfying the following:*

1. $|B_0||B_1| = 0 \Rightarrow |A_0||A_1| \geq 1$.
2. *For every $v \in V$ and $i = 0, 1$,*

$$v \in A_i \Rightarrow deg_{A_i}(v) - deg_{A_{1-i}}(v) \geq |deg_{B_i}(v) - deg_{B_{1-i}}(v)|.$$

3. *For every $v \in V$ and $i = 0, 1$,*

$$v \in B_i \Rightarrow deg_{B_{1-i}}(v) - deg_{B_i}(v) > |deg_{A_i}(v) - deg_{A_{1-i}}(v)|.$$

Proof. Suppose G is not an m.c.c. If G admits a consensus for every possible initial coloring c^0, there must exist d^0 for which G does not admit a majority consensus; i.e., there exists a coloring d^0 and t such that d^t is a consensus and $maj(d^0) \neq maj(d^t)$. Obviously, in the sequence $d^0, d^1, \ldots, d^t$, there exists $t' < t$ such that $maj(d^{t'}) \neq maj(d^{t'+1})$. Thus, (a) holds for $c^0 := d^{t'}$.

Thus, we may assume that there exists c^0 for which G does not admit a consensus. By Corollary 2.3 there exists t such that $c^t = c^{t+2}$. For $i = 0, 1$ define $A_i := \{v \in V : i = c_v^t = c_v^{t+1}\}$ and $B_i := \{v \in V : i = c_v^t \neq c_v^{t+1}\}$. Note that A_0, A_1, B_0, B_1 partition V and that item 1 must hold since neither c^t nor c^{t+1} is a consensus. Since for every $v \in A_i$, $c_v^t = c_v^{t+1}$,

$$deg_{A_i}(v) + deg_{B_i}(v) \geq deg_{A_{1-i}}(v) + deg_{B_{1-i}}(v).$$

Similarly, for every $v \in A_i$, $c_v^{t+1} = c_v^{t+2}$ implies (because $\{w : c_w^{t+1} = i\} = A_i \cup B_{1-i}$)

$$deg_{A_i}(v) + deg_{B_{1-i}}(v) \geq deg_{A_{1-i}}(v) + deg_{B_i}(v).$$

These two inequalities imply item 2. In the same manner, it follows that for every $v \in B_i$, $c_v^t \neq c_v^{t+1}$ implies

$$deg_{A_{1-i}}(v) + deg_{B_{1-i}}(v) > deg_{A_i}(v) + deg_{B_i}(v)$$

and that $c_v^{t+1} \neq c_v^{t+2}$ implies

$$deg_{A_i}(v) + deg_{B_{1-i}}(v) > deg_{A_{1-i}}(v) + deg_{B_i}(v).$$

Hence, item 3 follows from these two inequalities.

Conversely, suppose that (a) holds. Then, by Proposition 2.1, G is not an m.c.c.

Finally, suppose that (b) holds. Define $c_v^0 := i$ for $v \in A_i \cup B_i$, $i = 0, 1$. Note that item 3 implies that either none or both sets B_0 and B_1 must be nonempty (otherwise, if B_i is not empty and B_{1-i} is empty, the left-hand side of inequality in item 3 would be less than or equal to zero for a $v \in B_i$, thereby automatically violating the inequality). If both B_0 and B_1 are empty, item 1 implies that both A_0 and A_1 are not empty. Hence, in either case, c^0 is not a consensus because $A_i \cup B_i \neq \emptyset$ for $i = 0, 1$. Note that item 2 implies $c_v^1 = c_v^0$ for $v \in A_i$ and that item 3 implies that $c_v^1 \neq c_v^0$ for $v \in B_i$, $i = 0, 1$. Furthermore, item 2 also implies that $c_v^2 = c_v^1$ for $v \in A_i$, and item 3 implies that $c_v^2 \neq c_v^1$ for $v \in B_i$, $i = 0, 1$. Hence, $c^2 = c^0$ and the sequence $c^0, c^1, c^2, \ldots$ never admits a consensus. Thus, G is not an m.c.c. $\square$

Theorem 3.2 indicates possible ways of adding edges to G that is not an m.c.c. so that the new graph is still not an m.c.c. For example, if (b) holds for G and if there exists four edges defining a 4-cycle

$$E' := \{(w, x), (x, y), (y, z), (z, w)\}$$

such that $E' \cap E = \emptyset$, then it is straightforward to check that (b) holds for $G' = (V, E \cup E')$ provided that one of the following is true: (i) $w, x \in A_0$ and $y, z \in A_1$; (ii) $w, x \in B_0$ and $y, z \in B_1$; (iii) $w \in A_0$, $x \in B_0$, $y \in A_1$, and $z \in B_1$.

Special cases of Theorem 3.2 help identify large classes of graphs that are not m.c.c.'s and provide insight into the structure of graphs that are m.c.c.'s.

Corollary 3.3. *Let G be bipartite or disconnected. Then G is not an m.c.c.*

Proof. For a disconnected G, let $A_0 \neq V$ be one of G's connected components and set $A_1 = V \setminus A_0$, $B_0 = B_1 = \emptyset$. For a connected bipartite graph with bipartition $B_0 \cup B_1 = V$, set $A_0 = A_1 = \emptyset$. Note that items 1–3 from (b) in Theorem 3.2 hold. Thus, G is not an m.c.c. $\square$

Corollary 3.4. *Let G be an m.c.c.*

(a) *Every* min-*cut in G is trivial.*

(b) *G does not have a unique* max-*cut.*

Proof. First note that if (S, S^c) is a nontrivial min-cut, then for every $v \in S$, $deg_S(v) \geq deg_{S^c}(v)$, and for every $v \in S^c$, $deg_{S^c}(v) \geq deg_S(v)$ (if not, then (S, S^c) is not a min-cut since the cut $(S \cup \{v\}, S^c \setminus \{v\})$ has fewer edges). Thus, setting $A_0 = S$, $A_1 = S^c$, and $B_0 = B_1 = \emptyset$ gives a partition from Theorem 3.2(b). Hence, G is not an m.c.c.

Similarly, note that in a unique max-cut (S, S^c) for every $v \in S$, $deg_S(v) < deg_{S^c}(v)$, and for every $v \in S^c$, $deg_{S^c}(v) < deg_S(v)$ (if not, then (S, S^c) is not a unique max-cut since the cut $(S \cup \{v\}, S^c \setminus \{v\})$ has at least as many edges). Thus, setting $B_0 = S$, $B_1 = S^c$, and $A_0 = A_1 = \emptyset$ gives a partition from Theorem 3.2(b). Hence, G is not an m.c.c. $\square$

The last corollary indicates that m.c.c.'s are highly connected graphs (in the sense that having only trivial min-cuts and many max-cuts could be taken as a good indication of a high level of connectivity). The following theorem and its corollary provide another confirmation of this claim.

THEOREM 3.5. *Let G be an m.c.c. Then for every $v \in V$,*

$$\left| \bigcup_{w \in N(v) \cup \{v\}} N(w) \right| \geq (n+1)/2. \tag{3.1}$$

Proof. First, note that we can assume that G is connected (by Corollary 3.3) and that $n > 2$.

Suppose (3.1) does not hold for some $v \in V$. Let $u \in V$ be a vertex of the minimum degree among all vertices v for which (3.1) is violated. Let c^0 be a coloring such that $c^0_v = 1$ for every $v \in \bigcup_{w \in N(u) \cup \{u\}} N(w)$ and such that $sum(c^0) = (n+1)/2$. Note that $c^0_u = 1$ and that $maj(c^0) = 1$. Let d^0 be such that $d^0_v \neq c^0_v$ if and only if $v = u$ (i.e., the only difference between c^0 and d^0 is in the color of u). Note that $sum(d^0) = (n-1)/2$, and thus

$$maj(d^0) = 0 \neq 1 = maj(c^0). \tag{3.2}$$

Observe that for all $v \notin N(u) \cup \{u\}$, $w \in N(v)$ implies that $c^0_w = d^0_w$, and hence $c^1_v = d^1_v$. Further observe that $c^1_u = d^1_u = 1$ because the color of all neighbors of u is 1 in both c^0 and d^0 (and u has at least one neighbor since G is connected). Finally, observe that by the choice of u and the fact that G is connected and $n > 2$, $deg(v) \geq 2$ for all $v \in N(u)$. Since the color of all neighbors of v other than u is 1 in both c^0 and d^0, it follows that $c^1_v = d^1_v$ for $v \in N(u)$. Hence, $c^1 = d^1$ and thus, because of (3.2), either $maj(c^1) \neq maj(c^0)$ or $maj(d^1) \neq maj(d^0)$. In either case, it follows from Proposition 2.1 that G is not an m.c.c. □

The theorem shows that m.c.c.'s are nowhere truly local since the second neighborhood of any vertex contains a majority of the vertices of V. Hence, the local majority process always reaches a consensus on the initial majority color only if the local majority rule is *nowhere* local. Hence, the theorem can be viewed as a sort of impossibility result.

COROLLARY 3.6. *If G is an m.c.c., then $diam(G) \leq 4$ and $\Delta(G) \geq \lceil \sqrt{(n-1)/2}\, \rceil$.*

Proof. The proof follows immediately from (3.1) because for any two vertices $u, v \in V$,

$$\left(\bigcup_{w \in N(u) \cup \{u\}} N(w) \right) \bigcap \left(\bigcup_{w \in N(v) \cup \{v\}} N(w) \right) \neq \emptyset$$

and because $|\bigcup_{w \in N(v) \cup \{v\}} N(w)| \leq 1 + \Delta(G) + \Delta(G)(\Delta(G) - 1)$. □

We conjecture that a much stronger statement is true (this was confirmed to hold for $n \leq 13$ by an exhaustive search method).

MASTER CONJECTURE. *Every m.c.c. contains a master.*

This is a rather strong conjecture because it implies that a necessary condition for reaching majority consensus is the existence of a vertex connected to all the other vertices, thereby annihilating any notion of local computation. In the next section we'll show that the master conjecture holds for graphs G with $\delta(G) \geq n-3$. Note that, intuitively, such graphs should be considered as prime candidates for a counterexample to the conjecture since all of the vertices in these graphs are either masters or very close to being masters (i.e., 0-masters, 1-masters, or 2-masters). Thus, our result that the master conjecture holds for graphs with $\delta(G) \geq n-3$ provides strong evidence for the truth of the master conjecture.

4. The case of $\delta(G) \geq (n-3)$.

In this section, graphs with minimum degree $(n-3)$ are studied. We first show that every m.c.c. G with exactly $(n-3)/2$ master vertices has $\delta(G) \geq (n-3)$. (Note that determining whether a graph with at least $(n-1)/2$ masters is an m.c.c. is straightforward and does not depend on the degrees of the nonmaster vertices; cf. Proposition 3.1.) Then we turn to analysis of general graphs G with $\delta(G) \geq n-3$. We show that every such m.c.c. must have at least one master vertex; i.e., the master conjecture holds for G with $\delta(G) \geq n-3$. Furthermore, we give a complete characterization of m.c.c.'s with $\delta(G) \geq n-3$. We close the section by demonstrating that, for every n and positive $k \neq (n-1)/2$, there exists an m.c.c. whose number of masters is exactly k.

Theorem 4.1. *Let G be an m.c.c. with $(n-3)/2$ master vertices. Then $\delta(G) \geq (n-3)$.*

Proof. Let G be an m.c.c. with $(n-3)/2$ masters. Let M be the set of master vertices in G (so $|M| = (n-3)/2$). Let $N = V \backslash M$ and $K = \{v : (n-3) \leq deg(v) < (n-1)\}$.

Claim 1. $|K| \geq 2$.

Proof of Claim 1. Consider the partition $P_0 = M$ and $P_1 = V \setminus P_0$. For $v \in P_0$, set $c_v^0 = 0$, and for $v \in P_1$, set $c_v^0 = 1$. Note that $maj(c^0) = 1$. Since $P_0 = M$, by Proposition 2.6, for $v \in P_0$, $c_v^1 = maj(c^0) = 1$. Since G is an m.c.c., we have $maj(c^0) = maj(c^1)$. Hence, the set $\{v \in P_1 : c_v^1 = 1\}$ contains at least $(n+1)/2 - (n-3)/2 = 2$ vertices. We now show that $c_v^1 = 1$ if and only if $v \in K$. For a vertex $v \in P_1$, $c_v^1 = 1$ if and only if $deg_{P_0}(v) \leq deg_{P_1}(v)$. Since $deg_{P_0}(v) = |P_0| = (n-3)/2$, $deg_{P_1}(v) \geq (n-3)/2$. Therefore $deg(v) = deg_{P_0}(v) + deg_{P_1}(v) \geq (n-3)/2 + (n-3)/2 = (n-3)$, i.e., $v \in K$.

Claim 2. *For all $v \in N$, $deg_K(v) \leq |K| - 2$.*

Proof of Claim 2. For contradiction, assume there exists $v_i \in N$ such that $deg_K(v_i) > |K| - 2$. Then consider the partition $P_0 = M \cup \{v_i\}$, and $P_1 = V \setminus P_0$. Set $c_v^0 = 0$ for $v \in P_0$, and $c_v^0 = 1$ otherwise. For $v \in M$, $c_v^1 = maj(c^0) = 1$. Similarly, for all $v \in P_1 \setminus K$, $c_v^1 = 0$ since $deg(v) < n-3$ (by definition) and $deg_M(v) = (n-3)/2$. Now there are two cases.

(i) $v_i \in N \backslash K$. As $deg(v_i) < n-3$ (by definition) and $deg_M(v) = (n-3)/2$, we have $c_{v_i}^1 = 0$. For a vertex $v \in K$, $c_v^1 = 0$ if and only if $(v, v_i) \in E$ because $n-3 \leq deg(v) < n-1$ and $deg_{P_0}(v) = (n-1)/2$. Thus, $sum(c^1) = |M| + |K| - deg_K(v_i)$. Noting that $deg_K(v_i) > |K| - 2$, we have $sum(c^1) < (n-3)/2 + |K| - |K| + 2 = (n+1)/2$.

(ii) $v_i \in K$. Now $(n-3) \leq deg(v_i) < (n-1)$. Since $v_i \in K$, $deg_K(v_i) > |K| - 2$ implies that v_i is adjacent to all the other vertices in K, i.e., $deg_K(v_i) = |K| - 1$. For a vertex $v \in K \setminus v_i$, $c_v^1 = 0$ if and only if $(v, v_i) \in E$ since $deg(v) < n-1$ and $deg_{P_0}(v) = (n-1)/2$. Then $sum(c^1) = |M| + c_{v_i}^1 + (|K| - 1) - deg_K(v_i) \leq |M| + 1 + (|K| - 1) - (|K| - 1) = (n-1)/2$.

Thus $maj(c^1) = 0 \neq 1 = maj(c^0)$ in both (i) and (ii), which is a contradiction since G is an m.c.c.

Claim 3. *Let $v_1, v_2 \in K$ be vertices that are not adjacent. Then both v_1 and v_2 must be adjacent to all the vertices in $N \setminus K$.*

Proof of Claim 3. For contradiction, assume there exists a vertex $v_3 \in N \setminus K$ that is not adjacent to v_1. Note that $deg(v_1) = n-3$, and $deg_K(v_1) = |K| - 2$. Consider the partition $P_0 = M \cup v_1$, and $P_1 = V \setminus P_0$. Set $c_v^0 = 0$ if $v \in P_0$, and $c_v^0 = 1$ otherwise.

For $v \in M$, $c_v^1 = 1$. Also, note that $c_{v_1}^1 = 0$, since $deg_{P_0}(v_1) = (n-3)/2$, and $deg(v_1) = n-3$. Similarly, for $v \in K \setminus v_1$, $c_v^1 = 0$ if and only if $(v, v_1) \in E$. Therefore $sum(c^1) = |M| + (|K|-1) - deg_K(v_1) = (n-3)/2 + (|K|-1) - (|K|-2) = (n-1)/2$. This is a contradiction since G is an m.c.c.

We complete the proof of Theorem 4.1 by now showing that the size of the set $N \setminus K$ is zero; i.e., every vertex in N belongs to K. Any vertex $v \in K$ must have $deg_K(v) < |K|-1$ from Claim 2. Claim 3 implies that v is adjacent to all the vertices in $N \setminus K$, which implies that each vertex in $N \setminus K$ is adjacent to all the vertices in K. This is in contradiction to Claim 2, and therefore either the size of the set K is zero or the size of the set $N \setminus K$ is zero. From Claim 1, K is nonempty, and hence the set $N \setminus K$ must be empty. □

Now we turn our attention to graphs G with $\delta(G) \geq n-3$ and characterize m.c.c.'s. A direct consequence of Proposition 2.6 is that the only colorings c^0 for which G with $\delta(G) \geq n-3$ might not admit a majority consensus are the tight ones, i.e., c^0 such that $sum(c^0) = (n+1)/2$. (The case $sum(c^0) = (n-1)/2$ is symmetric.)

PROPOSITION 4.2. *If $\delta(G) \geq n-3$, then G admits majority consensus for every coloring c^0 such that $sum(c^0) \geq (n+3)/2$.*

Proof. Note that every $v \in V$ is either a master, a 1-master, or a 2-master. Thus, by Proposition 2.6, $c_v^1 = maj(c^0)$ for every $v \in V$. □

If $\delta(G) \geq n-3$, then G^c has a very simple structure since $\Delta(G^c) = (n-1) - \delta(G) \leq (n-1)-(n-3) = 2$. In other words, a connected component of G^c is a single vertex, a path, or a cycle. The decomposition of G^c into its connected components $H_1 = (V_1, E_1^c)$, $H_2 = (V_2, E_2^c)$, $\ldots$, $H_m = (V_m, E_m^c)$[4] will be used throughout this section and we will often abuse the notation and identify $V(H)$ with H whenever such notation is unambiguous (e.g., we will often say that the connected components of G^c define a partition of V).

Another convenient property of G with $\delta(G) \geq n-3$ is that every vertex in G is either a master, a 1-master, or a 2-master. Thus, the following lemma gives a complete Boolean formula representation of local updates for colorings c^t with $sum(c^t) = (n+1)/2$.

LEMMA 4.3. *Let c^t be a coloring such that $sum(c^t) = (n+1)/2$.*

(a) *If v is a master, then $c_v^{t+1} = 1$.*

(b) *If v is a 1-master, then $c_v^{t+1} = 1 - c_v^t c_w^t$, where w is the unique vertex not adjacent to v.*

(c) *If v is a 2-master, then $c_v^{t+1} = 1 - c_u^t c_w^t$, where u and w are the two vertices not adjacent to v.*

Proof. Since $maj(c^t) = 1$, (a) follows directly from Proposition 2.6.

If v is a 1-master, then $V \setminus N(v) = \{v, w\}$, so

$$|\{u \in N(v) : c_u^t = 1\}| = \frac{n+1}{2} - c_v^t - c_w^t.$$

Note that $c_v^{t+1} = 0$ if and only if $|\{u \in N(v) : c_u^t = 1\}| < |N(v)|/2 = (n-2)/2$ ($deg(v) = n-2$ is odd, so a tie is impossible). But the last inequality holds if and only if $c_v^t = c_w^t = 1$. Thus, (b) holds.

Similarly, if v is a 2-master, then $V \setminus N(v) = \{v, u, w\}$, so

$$|\{u \in N(v) : c_u^t = 1\}| = \frac{n+1}{2} - c_v^t - c_u^t - c_w^t.$$

[4] In other words, V_i, $i = 1, \ldots, m$, are pairwise disjoint, $V_1 \cup \cdots \cup V_m = V$, and $E_1^c \cup \cdots \cup E_m^c = E^c$.

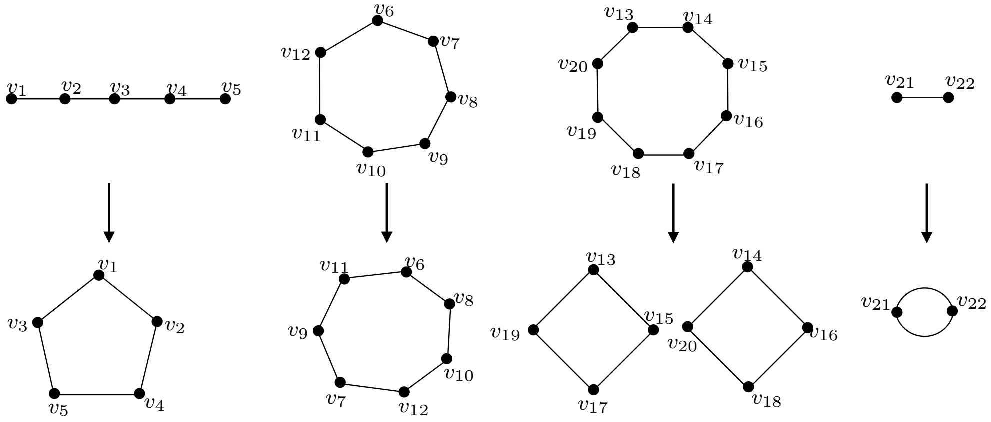

v_1 v_2 v_3 v_4 v_5
v_6 v_7 v_8 v_9 v_{10} v_{11} v_{12}
v_{13} v_{14} v_{15} v_{16} v_{17} v_{18} v_{19} v_{20}
v_{21} v_{22}
v_1 v_2 v_3 v_4 v_5
v_{11} v_6 v_8 v_9 v_{10} v_7 v_{12}
v_{13} v_{15} v_{19} v_{17} v_{14} v_{20} v_{16} v_{18}
v_{21} v_{22}

FIG. 4.1. *The auxiliary graphs (bottom) corresponding to the various connected components of G^c (top).*

First suppose $c_v^t = 0$. Then, $c_v^{t+1} = 0$ if and only if $|\{u \in N(v) : c_u^t = 1\}| \leq |N(v)|/2 = (n-3)/2$, and this is true if and only if $c_u^t = c_w^t = 1$. Thus, (c) holds if $c_v^t = 0$. Finally, suppose $c_v^t = 1$. Then, $c_v^{t+1} = 0$ if and only if $|\{u \in N(v) : c_u^t = 1\}| < |N(v)|/2 = (n-3)/2$ and, again, this is true if and only if $c_u^t = c_w^t = 1$. Thus, (c) also holds if $c_v^t = 1$. □

This lemma allows us to track the action of the local majority process on G. We define an *auxiliary graph* $AG = (V, E(AG))$. Edges of AG are defined by formulas from Lemma 4.3(b) and (c):

$$E(AG) \quad = \quad \{\{v,w\} : d_G(v) = n-2, \{v,w\} \notin E(G)\} \\ \cup \\ \{\{u,w\} : \exists v, d_G(v) = n-3, \{v,u\}, \{v,w\} \notin E(G)\}.$$

Thus, $E(AG)$ is in one-to-one correspondence with the set of all vertices of G which are not masters. Note that AG has a rather simple structure (see Figure 4.1 for an illustration): all of its connected components are cycles, each corresponding to a connected component of G^c as follows (this is a direct consequence of the definition of AG):

- If a connected component $H \subset G^c$ is a path, say $v_1, v_2, \ldots, v_l$ (i.e., $\{v_i, v_{i+1}\} \in E(G^c)$, $i = 1, \ldots, l-1$), then $V(H)$ defines a cycle C_H that is a connected component in AG.

  If l is even, then the adjacent vertices in C_H are

  $$v_1, v_2, v_4, v_6, \ldots, v_{l-2}, v_l, v_{l-1}, v_{l-3}, \ldots, v_3, v_1.$$

  (Note that if $l = 2$ for some H, AG becomes a multigraph, with C_H being a cycle of length 2.)

  If l is odd, then adjacent vertices in C_H are

  $$v_1, v_2, v_4, v_6, \ldots, v_{l-1}, v_l, v_{l-2}, v_{l-4}, \ldots, v_3, v_1.$$

- If a connected component $H \subset G^c$ is an odd cycle $v_1, v_2, \ldots, v_{2k+1}, v_1$ (i.e., $\{v_i, v_{i+1}\} \in E(G^c)$, $i = 1, \ldots, 2k+1$, and $\{v_{2k+1}, v_1\} \in E(G^c)$), then $V(H)$ defines a cycle C_H that is a connected component in AG:

  $$v_1, v_3, \ldots, v_{2k+1}, v_2, v_4, \ldots, v_{2k}, v_1.$$

- If a connected component $H \subset G^c$ is an even cycle $v_1, v_2, \ldots, v_{2k}, v_1$ (i.e., $\{v_i, v_{i+1}\} \in E(G^c)$, $i = 1, \ldots, 2k$, and $\{v_{2k}, v_1\} \in E(G^c)$), then $V(H)$ defines two disjoint cycles $C_H = C1_H \cup C2_H$ that are connected components in AG:

$$v_1, v_3, \ldots, v_{2k-1}, v_1 \text{ and } v_2, v_4, \ldots, v_{2k}, v_2.$$

LEMMA 4.4. *Let c^t be a coloring such that $sum(c^t) = (n+1)/2$. Let H be a connected component of G^c on l vertices, $l \geq 2$. Let $S = \{v \in H : c^t_v = 1\}$. Then*

$$|\{v \in H : c^{t+1}_v = 1\}| \geq l - |S|. \tag{4.1}$$

Furthermore, the equality holds in (4.1) *if and only if one of the following holds:* (i) $|S| = 0$, (ii) $|S| = l$, (iii) *H is an even cycle and $c^t_v \neq c^t_w$ whenever $\{v, w\} \in E(G^c)$.*

Proof. First note that, by Lemma 4.3 and by the definition of AG,

$$|\{v \in H : c^{t+1}_v = 1\}| = \sum_{v \in H} c^{t+1}_v = \sum_{\{u,w\} \in C_H \subset AG} (1 - c^t_u c^t_w) = |H| - \sum_{\{u,w\} \in C_H \subset AG} c^t_u c^t_w.$$

Thus, it remains to show that

$$|S| \geq \sum_{\{u,w\} \in C_H \subset AG} c^t_u c^t_w. \tag{4.2}$$

Note that

$$\sum_{\{u,w\} \in C_H \subset AG} c^t_u c^t_w = |\{\{u, w\} \in E_{AG}(C_H) : u, w \in S\}| = |E(C_H[S])|,$$

where $C_H[S]$ denotes the induced subgraph of C_H, i.e., the maximal subgraph of C_H on the vertex set $S \subset V(C_H)$.

If $|S| = 0$, $|E(C_H[S])| = 0$, and (4.2) holds with equality. Thus, (4.1) holds with equality.

If $|S| = l$, then $C_H[S] = C_H$ and $|E(C_H[S])| = |E(C_H)| = l$ since C_H is a cycle or a union of two disjoint cycles. Thus, if $|S| = l$, (4.1) also holds with equality.

If H is an even cycle, then $C_H = C1_H \cup C2_H$. Furthermore, $S = V(C1_H)$ or $S = V(C2_H)$ if and only if vertices of H are colored alternately along the cycle H (i.e., as described in (iii) in the statement of the lemma). In either case, $|E(C_H[S])| = |S|$ and (4.1) again holds with equality.

If neither (i) nor (ii) nor (iii) holds, then $C_H[S]$ contains an acyclic component, and any possible cyclic component of C_H must be a cycle.[5] Thus, $|E(C_H[S])| \leq |S| - 1$ and

$$|\{v \in H : c^{t+1}_v = 1\}| \geq l - |S| + 1. \qquad \square \tag{4.3}$$

Several simple consequences of this lemma will be useful in the analysis that follows. For example, if a connected component H of G^c that is not an isolated vertex is monochromatic for some c^t, then every vertex in H will switch color.

LEMMA 4.5. *Let c^t be a coloring of G, $\delta(G) \geq n-3$, such that $sum(c^t) = (n+1)/2$. Let $H = (V_H, E_H)$ be a connected component of G^c with $|V_H| \geq 2$. Suppose that $c^t_v = c^t_w$ for every $v, w \in V_H$. Then $c^{t+1}_v = 1 - c^t_v$ for every $v \in V_H$.*

[5] In fact, the only possibility for a cyclic component is when H is an even cycle.

Proof. Let $|V_H| = l$. If $c_v^t = 1$ for all $v \in V_H$, then the result follows from Lemma 4.4 with $|S| = l$. If $c_v^t = 0$ for all $v \in V_H$, then the result follows from Lemma 4.4 with $|S| = 0$. ∎

The next lemma presents an opposite scenario: If colors assigned by c^t, $sum(c^t) = (n+1)/2$, alternate along an even cycle that is a connected component of G^c, then no vertex on that cycle will switch color.

LEMMA 4.6. *Let c^t be a coloring of G, $\delta(G) \geq n-3$, such that $sum(c^t) = (n+1)/2$. Let $C_{2k} \subset G^c$ be a connected component in G^c. Suppose the colors assigned by c^t alternate along the cycle: If u is adjacent to v in C_{2k}, then $c_u^t = 1 - c_v^t$. Then $c_v^{t+1} = c_v^t$ for every $v \in C_{2k}$.*

Proof. Every $v \in C_{2k}$ is a 2-master and, by Lemma 4.3(c), $c_v^{t+1} = 1 - c_u^t c_w^t = 1 - (1 - c_v^t)$ because in C_{2k}, v is adjacent to both u and w. ∎

The preceding lemmas indicate a way to construct c^0 yielding a complete switch, i.e., $c^1 = 1 - c^0$. Obviously, all masters must be colored with a minority color in order to switch. If all the other connected components of G^c are monochromatic (with some even cycles possibly being colored as described in the previous lemma), and if the resulting coloring c^0 is a tight majority coloring on G (i.e., $sum(c^t) = (n+1)/2$), then, as shown in the next lemma, $c^1 = 1 - c^0$ (except on the even cycles, where $c^{t+1} = c^t$), and G is not an m.c.c.

LEMMA 4.7. *Let $\delta(G) \geq n-3$. Let $H_1 = (V_1, E_1^c)$, $H_2 = (V_2, E_2^c)$, …, $H_m = (V_m, E_m^c)$ be the connected components of G^c. Suppose there exist i and j, $1 \leq i < j \leq m$, such that*

(i) $|V_k| = 1 \Rightarrow k \leq i$,

(ii) $m \geq k > j \Rightarrow H_k$ *is an even cycle,*

(iii) $|V_1| + |V_2| + \cdots + |V_i| + 1 = |V_{i+1}| + \cdots + |V_j|$.

Then G is not an m.c.c.

Proof. For $v \in V_k$, set $c_v^0 = 0$ if $k \leq i$ and set $c_v^0 = 1$ if $i < k \leq j$. If $j < k \leq m$, then the remaining vertices lie on even cycles in G^c. Color each H_k alternately, i.e., as described in Lemma 4.6. Note that, by (iii),

$$\begin{aligned} |\{v \in V : c_v^0 = 0\}| &= \sum_{k=1}^{i} |V_k| + \frac{1}{2} \sum_{k=j+1}^{m} |V_k| \\ &= \left(\sum_{k=i+1}^{j} |V_k| \right) - 1 + \frac{1}{2} \sum_{k=j+1}^{m} |V_k| \\ &= |\{v \in V : c_v^0 = 1\}| - 1. \end{aligned}$$

Thus, $sum(c^0) = (n+1)/2$ and $maj(c^0) = 1$.

If v is a master, $c_v^1 = maj(c^0) = 1 = 1 - c_t^0$ (the last equality holds because $\{v\} = H_k$ for some k and $k \leq i$ by (i)). If v is not a master, then $v \in H_k$ for some $k \leq m$ such that $|H_k| \geq 2$. If $k \leq j$, then $c_v^1 = 1 - c_v^0$ by Lemma 4.5. If $k > j$, then $c_v^1 = c_v^0$ by Lemma 4.6. Therefore, $c_v^1 = 1 - c_v^0$ if $v \in V_1 \cup \cdots \cup V_j$ and $c_v^1 = c_v^0$ if $v \in V_{j+1} \cup \cdots \cup V_m$. Thus,

$$\begin{aligned} |\{v \in V : c_v^1 = 1\}| &= \sum_{k=1}^{i} |V_k| + \frac{1}{2} \sum_{k=j+1}^{m} |V_k| \\ &= \left(\sum_{k=i+1}^{j} |V_k| \right) - 1 + \frac{1}{2} \sum_{k=j+1}^{m} |V_k| \\ &= |\{v \in V : c_v^1 = 0\}| - 1. \end{aligned}$$

So, $maj(c^1) = 0 \neq maj(c^0)$ and G is not an m.c.c. by Proposition 2.1. ∎

For any $k = 0, 1, \dots, (n-1)/2$, it is straightforward to construct a G with k masters satisfying the conditions of Lemma 4.7. For example, if $k = 0$, take G such that connected components of G^c are $P_{(n-1)/2}$ and $P_{(n+1)/2}$. If $k > 0$, G such that connected components of G^c are masters in G, P_{k+1}, and C_{n-2k-1} is such an example. Thus, there exist G with $\delta(G) \geq n-3$, which are not m.c.c.'s, having exactly k masters for every $k < (n+1)/2$. (Recall that, by Proposition 3.1, every G with at least $(n+1)/2$ masters is an m.c.c.)

A similar construction to that of Lemma 4.7 yields a class of graphs with a unique master that are not m.c.c.'s.

LEMMA 4.8. *Let $\delta(G) \geq n-3$ and let v_0 be the unique master in G. Let $H_1 = \{v_0\}$, $H_2 = (V_2, E_2^c)$, …, $H_m = (V_m, E_m^c)$ be the connected components of G^c. Suppose there exist i and j, $1 \leq i \leq j \leq m$, such that*

(i) *$m \geq k > j \Rightarrow H_k$ is an even cycle,*

(ii) *$|V_2| + |V_3| + \cdots + |V_i| = |V_{i+1}| + \cdots + |V_j|$ (assuming the empty summation on both sides of the equation when $i = j = 1$).*

Then G is not an m.c.c.

Proof. Define c^0 as in the proof of Lemma 4.7 except for v_0. Set $c^0_{v_0} = 1$. Observe that $c^1_v = 1 - c^0_v$ for $v \in V_2 \cup \cdots \cup V_j$ (by Lemma 4.5), that $c^1_{v_0} = maj(c^0) = 1 = c^0_{v_0}$ (by Proposition 2.6), and that $c^1_v = c^0_v$ for $v \in V_{j+1} \cup \cdots \cup V_m$ (by Lemma 4.6). Note that $sum(c^1) = (n+1)/2$ because of (ii). Repeating the same observation, we get $c^2_v = 1 - c^1_v = 1 - (1 - c^0_v) = c^0_v$ for $v \in V_2 \cup \cdots \cup V_j$ (by Lemma 4.5), $c^2_{v_0} = maj(c^1) = 1 = c^0_{v_0}$ (by Proposition 2.6), and $c^2_v = c^1_v = c^0_v$ for $v \in V_{j+1} \cup \cdots \cup V_m$ (by Lemma 4.6). Thus $c^2 = c^0$, and $c^0, c^1, c^2, \dots$ is periodic with period at most two. Since c^0 is not a consensus, G is not an m.c.c. ∎

For example, $K_n \setminus C_{n-1}$ is not an m.c.c. because it satisfies the conditions of the lemma with $i = j = 1$ and $m = 2$.

In order to prove that the master conjecture holds in the case $\delta(G) \geq n-3$, we need yet another lemma. In what follows we will say that $v_1, v_2, \dots, v_k$ form a path P_k if v_i is adjacent to v_{i+1} for $i = 1, \dots, (k-1)$. Similarly, we will say that $v_1, \dots, v_k$ form a cycle C_k if $v_1, \dots, v_k$ form a path $P_k \subseteq C_k$ and v_1 is adjacent to v_k.

LEMMA 4.9. *Let c^t be a coloring of G, $\delta(G) \geq n-3$, such that $sum(c^t) = (n+1)/2$. Let $v_1, v_2, \dots, v_k$ form $H \subset G^c$, a connected component in G^c on $k \geq 3$ vertices. Suppose that there exists a $j < k/2$ such that $c^t_{v_i} = i \mod 2$ for $i \leq 2j+1$. If $2j+1 < k$, also suppose that $c^t_{v_i} = c^t_{v_{2j+2}}$ for $i > 2j+1$.*

Then $c^{t+1}_{v_i} = c^t_{v_i}$ for $i \leq 2j+1$ and $c^{t+1}_{v_i} = 1 - c^t_{v_i}$ for $i > 2j+1$.

Proof. Since $\delta(G) \geq n-3$, H is a path or a cycle. Using Lemma 4.3(b) and (c), observe that $c^{t+1}_{v_i} = c^t_{v_i}$ for $i \leq 2j+1$ (since each v_i such that $c^t_{v_i} = 0$ has both nonneighbors of color 1, while each v_i such that $c^t_{v_i} = 1$ has at least one nonneighbor of color 0) and that $c^{t+1}_{v_i} = 1 - c^t_{v_i}$ for $i > 2j+1$ (if $c^t_{v_{2j+2}} = \cdots = c^t_{v_k} = 0$, then each such v_i has a nonneighbor of color 0; if $c^t_{v_{2j+2}} = \cdots = c^t_{v_k} = 1$, then each such v_i has all nonneighbors of color 1 because $c^t_{v_1} = c^t_{v_{2j+1}} = 1$). ∎

THEOREM 4.10. *Let G be a graph such that $\delta(G) \geq n-3$. If G is an m.c.c., then G contains a master.*

Proof. Suppose G does not contain a master. We'll show that G is not an m.c.c. Let $H_1 = (V_1, E_1^c)$, $H_2 = (V_2, E_2^c)$, …, $H_m = (V_m, E_m^c)$ be the connected components of G^c. Since G does not contain a master, $|V_l| \geq 2$, $l = 1, \dots, m$. Choose an index i such that

$$|V_1| + \cdots + |V_i| \leq (n-1)/2 < |V_1| + \cdots + |V_i| + |V_{i+1}|.$$

If $|V_1|+\cdots+|V_i| = (n-1)/2$, then the conditions of Lemma 4.7 are satisfied with i and with $j = m$. Therefore, in this case, G is not an m.c.c.

For the rest of the proof we may assume that $|V_1| + \cdots + |V_i| < (n-1)/2$. We may also assume that $|V_1| + \cdots + |V_{i-1}| + |V_i| + (|V_{i+1}|/2) > (n-1)/2$. (If not, then $(|V_{i+1}|/2) + |V_{i+2}| + |V_{i+3}| + \cdots + (|V_m|) > (n-1)/2$ and we could map l to $m+1-l$; i.e., H_l becomes H_{m+1-l}, $l = 1, \dots, m$.) Note that these imply that $|V_{i+1}| \geq 3$.

Let $v_1, v_2, \dots, v_k$ form H_{i+1} and let

$$j = (n-1)/2 - (|V_1| + \cdots + |V_{i-1}| + |V_i|). \tag{4.4}$$

Note that $j < k/2$. Set

$$c_v^0 = \begin{cases} \quad 0 & v \in V_1 \cup V_2 \cup \cdots \cup V_i, \\ p \mod 2 & v_p, \quad p = 1, \dots, 2j+1, \\ \quad 1 & v_p, \quad p = 2j+2, \dots, k, \\ \quad 1 & v \in V_{i+2} \cup V_{i+3} \cup \cdots \cup V_m. \end{cases}$$

Note that $sum(c^0) = (n+1)/2$. By Lemma 4.5, $c_v^1 = 1 - c_v$ for every $v \notin V_{i+1}$. By Lemma 4.9, $c_{v_i}^1 = 1 - c_{v_i}^0$ for $i = 2j+2, \dots, k$ and $c_{v_i}^1 = c_{v_i}^0$ for $i = 1, \dots, 2j+1$. Thus, only j vertices colored by 0 and only $j+1$ vertices colored by 1 do not switch color. Hence, $sum(c^1) = |V_1| + \cdots + |V_i| + (j+1) = (n-1)/2 + 1 = (n+1)/2$ (the second equality follows from (4.4)).

Repeating the same argument for

$$c_v^1 = \begin{cases} \quad 1 & v \in V_1 \cup V_2 \cup \cdots \cup V_i, \\ p \mod 2 & v_p, \quad p = 1, \dots, 2j+1, \\ \quad 0 & v_p, \quad p = 2j+2, \dots, k, \\ \quad 0 & v \in V_{i+2} \cup V_{i+3} \cup \cdots \cup V_m, \end{cases}$$

we conclude that $c^2 = c^0$. Thus, $c^0, c^1, c^2, \dots$ has period two. Therefore, G is not an m.c.c. □

Next we turn to G, $\delta(G) \geq n-3$, which contain masters. Because of Proposition 3.1, the only remaining cases are graphs with k masters, $k = 1, 2, \dots, (n-3)/2$. We have already demonstrated two conditions that would immediately classify such G as not being an m.c.c. (Lemmas 4.7 and 4.8). As the next theorem shows, these are the only two obstacles.

THEOREM 4.11. *Let G, $\delta(G) \geq n-3$, contain exactly k masters, $1 \leq k \leq (n-3)/2$. G is not an m.c.c. if and only if G satisfies conditions of Lemma* 4.7 *or conditions of Lemma* 4.8.

Proof. We only have to prove necessity. (Sufficiency follows from Lemmas 4.7 and 4.8.) We will show that, for any c^0, G either admits a majority consensus for c^0 or satisfies conditions of either Lemma 4.7 or Lemma 4.8. By Proposition 4.2 we may assume $sum(c^0) = (n+1)/2$.

Let $H_1, H_2, \dots, H_l$ be the connected components of G^c that are not isolated vertices. Let c^0 be a coloring of G, $sum(c^0) = (n+1)/2$. For $i = 0, 1$ define

$$m(i) = |\{v \in G : v \text{ is a master, } c_v^0 = i\}|$$

and

$$h_j(i) = |\{v \in H_j : c_v^0 = i\}|,$$

$j = 1, \dots, l$. Note that $k = m(0) + m(1)$, $|H_j| = h_j(0) + h_j(1)$, $(n+1)/2 = sum(c^0) = m(1) + \sum_{j=1}^{l} h_j(1)$ and that $(n-1)/2 = m(0) + \sum_{j=1}^{l} h_j(0)$.

Furthermore, for $j = 1, \dots, l$, let $\alpha_j = 0$ if H_j satisfies (i), (ii), or (iii) from Lemma 4.4; otherwise let $\alpha_j = 1$. In this notation, by Lemma 4.4, we have

$$|\{v \in H_j : c_v^1 = 1\}| \geq |H_j| - h_j(1) + \alpha_j = h_j(0) + \alpha_j.$$

Thus, taking into account that $c_v^1 = 1$ for every master v (Proposition 2.6),

$$sum(c^1) \geq m(0) + m(1) + \sum_{j=1}^{l} (h_j(0) + \alpha_j).$$

Thus,

$$\begin{aligned} sum(c^1) - sum(c^0) \quad &\geq m(1) + m(0) + \sum_{j=1}^{l} h_j(0) + \sum_{j=1}^{l} \alpha_j - sum(c^0) \\ &\geq m(1) + (n-1)/2 + \sum_{j=1}^{l} \alpha_j - (n+1)/2 \\ &\geq m(1) - 1 + \sum_{j=1}^{l} \alpha_j. \end{aligned}$$

Therefore, $sum(c^1) < sum(c^0)$ if and only if $m(1) = 0$ and $\alpha_j = 0$ for all $j = 1, \dots, n$. Also, in this case, $sum(c^1) = (n-1)/2$; i.e., G is not an m.c.c. since $maj(c^1) = 0 \neq 1 = maj(c^0)$. Note that conditions of Lemma 4.7 are satisfied; list all components of G^c as follows: Start with all masters and continue with all H such that $c_v^0 = 0$ for all $v \in H$; then continue by listing all H such that $c_v^0 = 1$ for all $v \in H$; if there are remaining components, these must be even cycles colored alternately along the cycle by c^0.

If $sum(c^1) > sum(c^0)$, then $sum(c^1) \geq (n+3)/2$ and, therefore, G admits majority consensus for c^1 by Proposition 4.2 (and thus for c^0 also since $maj(c^0) = maj(c^1)$).

Therefore, we may assume that $sum(c^1) = sum(c^0) = (n+1)/2$. This also means that we may assume that

$$1 - m(1) = \sum_{j=1}^{l} \alpha_j.$$

Since the right-hand side is nonnegative, $m(1) = 0$ or $m(1) = 1$. Note that we may assume that $m(1) = 1$. (Otherwise, we can replace c^0 with c^1 and, since $sum(c^1) = (n+1)/2$ with α_j unchanged, we can ensure that $m(1) \neq 0$.)

Because $m(1) = 1$, we know that every $\alpha_j = 0$; i.e., H_j either is colored monochromatically by c^0 or is an even cycle colored alternately by c^0. If, in addition, $m(0) = 0$, G has only one master and the conditions of Lemma 4.8 are satisfied; list all components of G^c as follows: Start with the unique master, continue with all H such that $c_v^0 = 0$ for every $v \in H$, and then continue by listing all H such that $c_v^0 = 1$ for every $v \in H$; if there are any vertices in H left, these must be even cycles colored alternately.

Therefore, it remains to consider the case $sum(c^0) = sum(c^1) = (n+1)/2$, $m(1) = 1$, $m(0) \geq 1$, and $\alpha_j = 0$ for every H_j. Note that

$$\begin{aligned} sum(c^2) &= m(0) + m(1) + \sum_{j=1}^{l} h_j(1) \geq 1 + m(1) + \sum_{j=1}^{l} h_j(1) \\ &= 1 + sum(c^0) = (n+3)/2. \end{aligned}$$

Thus, by Proposition 4.2, c^3 is a consensus in color 1. □

Remark. The proof of Theorem 4.11 shows that if G, with $\delta(G) \geq n-3$, is an m.c.c., then c^3 is the majority consensus; i.e., the local majority process reaches consensus in at most three steps.

THEOREM 4.12. *Let G be a graph with k masters and with $\delta(G) \geq n-3$. G is not an m.c.c. if and only if one of the following holds:* (i) $k = 0$, (ii) $k = (n-1)/2$, (iii) *G satisfies the conditions of Lemma* 4.7, (iv) *G satisfies the conditions of Lemma* 4.8.

Proof. The proof follows by Theorem 4.10, Proposition 2.6, and Theorem 4.11. □

Note that Theorem 4.11 can be used to define various classes of m.c.c.'s and non-m.c.c.'s. We close this section by observing just one additional class of m.c.c's that was mentioned in section 2.

THEOREM 4.13. *Let $k \geq 1$, $k \neq (n-1)/2$. There exists an m.c.c. with exactly k masters. In particular, $K_n \setminus P_{n-k}$ is an m.c.c.*

Proof. If $k \geq (n+1)/2$, the result follows from Proposition 3.1. If $1 \leq k \leq (n-3)/2$, the connected components of the complement of $G = K_n \setminus P_{n-k}$ are k single element components corresponding to masters in G and P_{n-k}. Note that G does not satisfy the conditions of Lemma 4.7 or Lemma 4.8. Thus, by Theorem 4.11, G is an m.c.c. □

Finally, checking if the conditions of Lemmas 4.7 and 4.8 are satisfied can be done in polynomial time. This is because the exact structure of connected components of G^c can be found in polynomial time and because the (nearly) equipartition conditions can be checked in time polynomial in $n = \sum |V_i|$ (this is essentially a knapsack problem that can be solved in pseudopolynomial time [13]). Thus, in light of Theorem 4.12, we have the following.

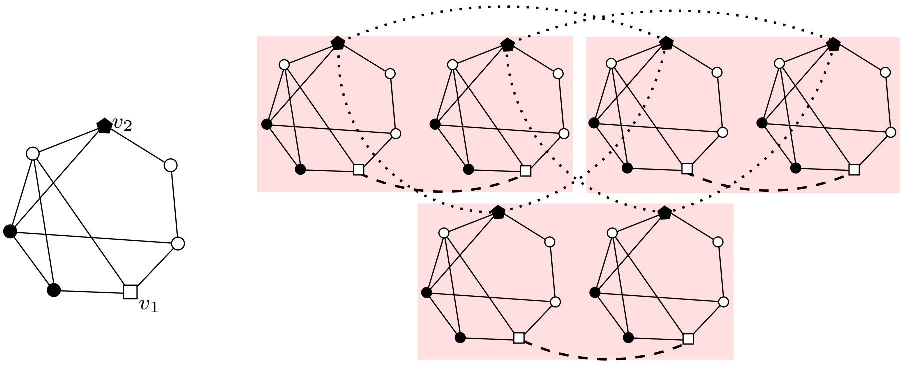
COROLLARY 4.14. *Let G be a graph with $\delta(G) \geq n-3$. Then determining whether G is an m.c.c. can be done in polynomial time.*

5. Generalizations and relaxations.

A simple generalization of the local majority process would allow vertex v to have some resistivity towards color switch. Formally, for a nonnegative integer $k(v)$, we define a $k(v)$-*local majority rule* for vertex v:

$$c_v^{t+1} = \begin{cases} c_v^t & \text{if } |\{w \in N_v : c_w^t = c_v^t\}| \geq |N(v)|/2 - k(v), \\ 1 - c_v^t & \text{if } |\{w \in N_v : c_w^t \neq c_v^t\}| > |N(v)|/2 + k(v). \end{cases} \tag{5.1}$$

The value $k(v)$ is called the *resistivity value* of vertex v and we call the graph $G = (V, E)$, together with the set of vertex resistivities $\{k(v) : v \in V\}$, a *varied-resistivity* graph. Similarly, the process defined by (5.1) is called the local majority process with resistivities. Dreyer [7] studies such processes and discusses relevant literature. Note that the local majority process with resistivities, where $k(v) = 0$, $v \in V$, is exactly the local majority process.



G

$\tilde{G}$

(a) (b)

FIG. 5.1. (a) *Graph G with $k(v_1) = 1$, $k(v_2) = 2$, and c^0 as indicated.* (b) *$\tilde{G}$ with $\tilde{c}^0$ as indicated.*

As the following theorem shows, introducing vertex resistivities does not introduce additional difficulties: For any coloring c^0 and any graph G with resistivities $\{k(v) : v \in V\}$, the local majority process with resistivities can be simulated by the (standard) local majority process on a related graph $\tilde{G}$ containing $\prod_v (k(v)+1)$ disjoint copies of G that are interconnected through $\prod_{v \neq v_i}(k(v)+1)$ vertex-disjoint $(k(v_i)+1)$-cliques on vertices corresponding to v_i in each of these copies for every v_i and with the initial coloring $\tilde{c}^0$ on $\tilde{G}$ coinciding with c^0 on each of the copies of v_i. (Figure 5.1 provides an example.)

THEOREM 5.1. *Let $G(V, E)$ be a varied-resistivity graph with vertex set $V = \{v_1, \dots, v_n\}$ and corresponding resistivities $\{k(v_1), \dots, k(v_n)\}$. The local majority process with resistivities on the varied-resistivity graph G can be simulated by the local majority process on some graph $\tilde{G}(\tilde{V}, \tilde{E})$.*

Proof. Denoting $\mathbf{p} = (p_1, p_2, \dots, p_n)$ and $\mathbf{q} = (q_1, q_2, \dots, q_n)$, define $\tilde{G} = (\tilde{V}, \tilde{E})$ as follows:

$$\begin{aligned} \tilde{V} &= \{v_i^{\mathbf{p}} : v_i \in V, 0 \leq p_l \leq k(v_l), i = 1, \dots, n, l = 1, \dots, n\}, \\ \tilde{E} &= \{\{v_i^{\mathbf{p}}, v_j^{\mathbf{p}}\} : \{v_i, v_j\} \in E, 0 \leq p_l \leq k(v_l), l = 1, \dots, n\} \\ &\quad \cup \ \{\{v_i^{\mathbf{p}}, v_i^{\mathbf{q}}\} : p_l = q_l \Leftrightarrow l \neq i, i = 1, \dots, n\}. \end{aligned}$$

Note that, for any $\mathbf{p} = (p_1, \dots, p_i, \dots, p_n)$, the subgraph of $\tilde{G}$ induced by $\{v_i^{\mathbf{p}} : v_i \in V\}$ is isomorphic to G, and that, for any i, the subgraph of $\tilde{G}$ induced by $\{v_i^{\mathbf{p}} : 0 \leq p_i \leq k(v_i)\}$ is a $(k(v_i)+1)$-clique. Furthermore, the edge sets of these $\prod_i (k(v_i)+1) + \sum_i \prod_{j \neq i}(k(v_j)+1)$ subgraphs partition $\tilde{E}$. Thus,

$$N(v_i^{\mathbf{p}}) \ = \ \{v_j^{\mathbf{p}} : v_j \in N(v_i)\} \ \cup \ \{v_i^{\mathbf{q}} : p_l = q_l \Leftrightarrow l \neq i\};$$

that is, $N(v_i^{\mathbf{p}})$ consists of the set of vertices isomorphic to $N(v_i)$ and $k(v_i)$ additional vertices.

In order to simulate the local majority process with resistivities on G by the local majority process (without resistivities) on $\tilde{G}$, for any coloring c^t of G, define $\tilde{c}^t$ by

$$\tilde{c}^t(v_i^{\mathbf{p}}) = c^t(v_i).$$

Since $\tilde{c}^t(v_i^{\mathbf{p}}) = \tilde{c}^t(v_i^{\mathbf{q}})$ for all $\mathbf{p}, \mathbf{q}$, $\tilde{c}^{t+1}(v_i^{\mathbf{p}}) = 1 - \tilde{c}^t(v_i^{\mathbf{p}})$ if and only if $|\{v_j^{\mathbf{p}} : v_j \in N(v_i), \tilde{c}^t(v_j^{\mathbf{p}}) = 1 - \tilde{c}^t(v_i^{\mathbf{p}})\}| > |\{v_j^{\mathbf{p}} : v_j \in N(v_i)\}|/2 + k(v_i) = |N(v_i)|/2 + k(v_i)$. Thus, using (5.1),

$$\tilde{c}^{t+1}(v_i^{\mathbf{p}}) = c^{t+1}(v_i),$$

that is, $c^{t'}$ on G and $\tilde{c}^{t'}$ on any subgraph of $\tilde{G}$ induced by $\{v_i^{\mathbf{p}} : v_i \in V\}$ coincide for any $t' \geq t$. □

Another natural generalization of our model would be to associate weights to the edges of G. The weight a_{uv} associated to the edge $\{u, v\}$ could be interpreted as the strength of the relationship between u and v and used as the relative importance of the information provided by $u \in N(v)$ at time t for v's decision on its color c_v^{t+1} at time $t+1$. In other words, the *weighted local majority process* is defined by

$$c_v^{t+1} = 1 - c_v^t \Leftrightarrow \left(\sum_{u \in N(v): c_u^t = 1 - c_v^t} a_{uv} \right) > \frac{1}{2} \sum_{u \in N(v)} a_{uv}.$$

Note that we may assume that all a_{uv} are rational numbers because G is finite.[6] Furthermore, we may assume that all weights a_{uv} are integer valued.[7] If all weights happen to be nonnegative, the weighted local majority process can be simulated by the local majority process on the corresponding multigraph $G^M = (V, E^M)$, where E^M is the multiset of edges of G with multiplicity of an edge $\{u, v\}$ given by a_{uv}. In order to properly apply (2.1) in this case, several definitions have to be adjusted: $N(v)$ is the multiset of the vertices adjacent to v, where a vertex u appears with multiplicity a_{uv}. A degree of a vertex v, $deg(v)$ is the cardinality (taking into account multiplicities) of $N(v)$. Most of our results readily generalize in the multigraph framework.

Combining both generalizations yields the *local weighted majority process with resistivities* which is defined by

$$c_v^{t+1} = 1 - c_v^t \Leftrightarrow \left(\sum_{u \in N(v): c_u^t = 1 - c_v^t} a_{uv} \right) > k_v + \left(\sum_{u \in N(v): c_u^t = c_v^t} a_{uv} \right). \tag{5.2}$$

In other words, v changes its color at time $t+1$ if and only if at least k_v more than the weighted majority of its neighbors are colored by $1 - c_v^t$ at time t. As already noted, it is safe to assume that edge weights a_{uv} are integral, and thus, without loss of generality, we may assume that resistivities k_v are integral also.[8] If k_v is nonnegative, this process can be simulated by the local majority process on the multigraph G^M defined as in the previous paragraph with addition of k_v loops to each vertex v. Note that v belongs to $N(v)$ with multiplicity k_v in such a multigraph.

Hence, because of the multigraph simulation, it is straightforward to generalize most of the presented results in the framework that allows for nonnegative edge weights a_{uv} and nonnegative vertex weights k_v. Note that the weighted local majority process with resistivities can be described in terms of the process from Theorem 2.2 with $b = \frac{1}{2}A(1, \ldots, 1)^T$, where nondiagonal entries of the integer valued A are a_{uv} if

[6] There exists $\epsilon > 0$ such that replacing each weight a_{uv} with a_{uv}^*, $a_{uv} - \epsilon < a_{uv}^* \leq a_{uv}$ yields the same process for any initial coloring c^0.

[7] Multiplying all weights by a scalar λ yields the same process. Set λ to the common denominator of all weights.

[8] Any k_v can be replaced with an integral k_v^* resulting in no change in the coloring process.

$\{u, v\}$ is an edge and $a_{uv} = 0$ if $\{u, v\}$ is not an edge, and where the diagonal entries of A are $a_{vv} = k_v + 1$ if there exists c^t which would turn the right-hand side of (5.2) into equality, and $a_{vv} = k_v$ otherwise. Conversely, if A is a nonnegative matrix and b such that, for every $i = 1, \dots, n$, the interval $(b_i - a_{ii}, b_i)$ contains no elements from

$$S = \left\{ \sum_{j \in J} a_{ij} : J \subseteq \{1, 2 \dots, n\} \setminus \{i\} \right\}$$

and $b_i \geq (\max S)/2$, then it is straightforward to define G and nonnegative edge and vertex weights such that the weighted local majority process with resistivities on G is equivalent to the dynamic process in the symmetric neural network model (i.e., the process described in Theorem 2.2). Finally, note that there are instances of the general symmetric neural network process, e.g., a weighted local majority process with resistivities, where any of the edge weights or vertex resistivities are negative, that cannot be simulated by the multigraph simulation approach to the local majority process. (Allowing negative weights and resistivities could be viewed as a technical generalization but not necessarily a natural one. If at least one of these parameters is negative, there exists v and a coloring c^0 such that v switches its color at time $t = 1$ from c_v^0, which is the majority color of its neighborhood at time $t = 0$, to $c_v^1 = 1 - c_v^0$, which is the minority color of its neighborhood at time $t = 0$. A rule that allows for a switch from the local majority to the local minority hardly qualifies as an acceptable majority computation rule.)

In the next section we further discuss some basic assumptions of our model and try to illustrate why our model is a natural one to analyze. In the rest of this section we present an approach towards relaxing the notion of m.c.c.

In view of our results showing that m.c.c.'s are nowhere truly local, one might want to know how likely is a network of agents G to admit a majority consensus. Here we single out and then combine two possible ways to measure this. On the one hand, one might be interested only in the colorings, where the difference between majority and minority is substantial; i.e., we may only require that G admit majority consensus only for a coloring c^0 such that $sum(c^0) \geq k$ (where an integer $k \geq n/2$ is a numerical expression of "substantial majority"). On the other hand, one might allow for occasional failures requiring that G admit majority consensus for a substantial proportion p of initial colorings c^0 that are of interest. These two approaches motivate the following definition of the (p, k)-weak m.c.c. defined for $0 \leq p \leq 1$ and integer $k > n/2$. G is a *(p, k)-weak m.c.c.* if G admits a majority consensus for at least $p|C(k)|$ colorings $c^0 \in C(k) = \{c^0 : sum(c^0) \geq k\}$. Note that, for odd n, m.c.c. is equivalent to a $(1, (n + 1)/2)$-weak m.c.c. Also note that, if G is a (p, k)-weak m.c.c. and a (p', k')-weak m.c.c., then $k \leq k' \Rightarrow p \leq p'$. (This follows from Lemma 2.4.) Further note that, by Proposition 2.6, any G with $\delta(G) \geq 2(n - k)$ is a $(1, k)$-weak m.c.c. In fact, we believe that the following generalization of the master conjecture holds.

GENERALIZED MASTER CONJECTURE. *Every $(1, k)$-weak m.c.c. contains a $(2k - n - 1)$-master.*

6. Discussion of model assumptions. As already noted, one might think that our model is neither a realistic one nor a natural one to study because of several assumptions that we have made. In this section we discuss model assumptions and hopefully illustrate why our model is a natural one to study.

Choice of the neighborhood and the tie-breaking rule. One might consider our choice of the tie-breaking rule and the definition of the neighborhood somewhat arbitrary. For example, why not redefine $N(v)$ by including v itself in $N(v)$ and/or modify the tie-breaking rule so that $c_v^{t+1} = 1 - c_v^t$ whenever $|\{w \in N(v) : c_w^t = 1\}| = |N(v)|/2$ (i.e., require that v switch color in the case of a tie in the neighborhood). Note that adopting both changes simultaneously in our model does not change the process. Also note that the proposed modification of the tie-breaking rule (without redefining $N(v)$) allows no m.c.c.'s.[9] Other modifications can be found in the literature (e.g., several variants are studied in [35]).

Since it is impossible to discuss all possible creative proposals for the modifications of our model, let's discuss some properties that a reasonable model should have, and then show that our model is the only one satisfying these properties.

The least one should expect from a local majority process is that every vertex v should be able to update its color c_v^t so that c_v^{t+1} is the majority color among the colors it is aware of; i.e., c_v^{t+1} should be the majority color on $N(v) \cup \{v\}$ at time t. In the case where the majority is not defined, the update should reflect that ambivalence; i.e., if there is a tie among colors that v is aware of at time t, then $c_v^{t+1} = 1 - c_v^t$. In other words, c_v^{t+1} should be computed to reflect the majority situation in $N(v) \cup \{v\}$ at time t (majority is 0 or 1 or ambivalent) because v has no information about the possible existence of vertices not in $N(v) \cup v$. Thus, if v happens to be a master, thereby having no reasons for faulty computations of global majority, c_v^{t+1} will correctly signal the global majority.[10] Note that the stated properties uniquely define

$$c_v^{t+1} = f(\{c_w^{t+1} : w \in N(v) \cup \{v\}\}),$$

and (2.1) is a way to represent c^{t+1}. Therefore, if the goal is to define a local update step satisfying outlined properties, the only choice is the local update used in our model.

Bidirectional communication. The bidirectional nature of the relationship among the agents played a crucial rule in our analysis. For example, even the basic *"period is one or two"* property does not hold when G is allowed to be a directed graph. For example, if $\vec{C_n}$ is a directed cycle on n vertices and c^0 is the coloring assigning 1 to only one vertex and 0 to the remaining $n-1$ vertices, $c^0, c^1, c^2, \ldots$ is periodic with period n. Thus, allowing for nonsymmetric relationships yields to periods of any possible length. In order to generalize presented results, one would have to take into account the possibility of periods longer than two.

Memoryless property. The memoryless property of the local majority process might seem unreasonable in many applications. In this paper we investigated iterative use of the local majority rule as the simplest local approach to the problem of determining global majority. Limited computational power of the agents due to the memoryless property of the process and the agent's ability to calculate and communicate the local majority in the form of one-bit information is of central importance in our analysis. Empowering agents with memory would bring the problem closer to the standard distributed computing framework. Design and analysis of possibly more successful and more complicated protocols of the distributive computing flavor is beyond the scope of this paper. Here we only note that the problem of determining majority becomes trivial if all agents are aware of the network structure. (If G is disconnected, there is no way to communicate between two connected components.

[9] Let $c_v^0 = 1$ if and only if $v \in S$, where (S, S^c) is a max-cut in G. If the tie-breaking rule is redefined as described, then $c^1 = 1 - c^0$ (because $deg_{S^c}(v) \geq deg_S(v)$ for every $v \in S$).

[10] If, for a master v, $c_v^{t+1} = 1 - c_v^t$, one has to check c_v^{t+2} to determine if the color switch indicated ambivalence or the choice of the global majority color.

If G is connected, the information about c_v^0 can be propagated through the network. This could be repeated for all n vertices which would allow all agents to learn c^0 and thus $maj(c^0)$). Thus, the interesting protocols would be those defined for agents that have no unique identifications and have no information about the network.

Static network structure. The static nature of the network of agents is another critical property of the local majority process. It is possible that allowing for network dynamics in the form of protocols that simultaneously control changes in c^t and the structure of the network at time t (e.g., changing the weight of an edge; adding/deleting an edge) might yield efficient protocols. This seems to be a fundamentally different model than the one studied here.

Synchronous versus asynchronous updates. Synchronous updates make the local majority process less restrictive than it would be with possible asynchronous update protocols. If the local majority process is modified in a way that an infinite sequence $v_1, v_2, \ldots$ of vertices from G is given and that the only update of c^t at time t occurs at vertex v_t according to the local majority rule (2.1) while $c_v^{t+1} = c_v^t$ for all $v \neq v_t$, then no G except the complete graph on an odd number of vertices can be an m.c.c. (First, note that v_1 must be a master with $deg(v) = n - 1$ even to ensure that $maj(c^0) = maj(c^1)$ for all colorings c^0. Thus, by induction, all vertices appearing in the sequence must be masters with even degree. If a vertex v does not appear in the sequence, then G cannot be an m.c.c. because an update at v will never occur.)

Deterministic versus stochastic model. The presented model is purely deterministic and there are several aspects of the model that call for stochastic modification. For example, it would be interesting to see the effect of replacing the local update (2.1) by the stochastic update rule

$$\mathtt{P}(c_v^{t+1} = i) = \frac{1}{deg(v)} |\{w \in N(v) : c_w^t = i\}|$$

on the conclusions drawn from the model. (The same rule is used in [21, 22, 32, 33].)

Also, allowing for asynchronous updates where the next vertex to be updated is selected at random could yield interesting results. However, one has to be aware that stochastic rules allow for a nonzero probability of not admitting a majority consensus.

Number of colors. One might consider generalizing the model by allowing k possible colors, i.e., allowing $c^t : V \to \{0, 1, \ldots, k-1\}$. Properly defining the tie-breaking rule is an inherent problem of this generalization. If $k > 2$, it is possible that c_v^t is a minority color in $N(v) \cup \{v\}$ and that there is more than one majority color in $N(v)$. Then any tie-breaking choice for c_v^{t+1} would have to favor one of the majority colors arbitrarily. Regardless of the definition of the tie-breaking rule that would hopefully generalize the one used in our model, understanding the case $k = 2$ is a prerequisite for understanding models allowing $k > 2$ colors. (Clearly, if G is an m.c.c. when k colors are possibly present, G is also an m.c.c. when $k' < k$ are possibly present.)

A generalization that would be more along the lines of our approach would be to allow $c^t : V \to \mathbf{R}$, define the dynamic process by

$$c_v^{t+1} = \sum_{w \in N(v)} c_w^t,$$

and state that G admits a majority consensus for coloring c^0 if there exists a t such that $sign(c_v^t) = sign(sum(c^0))$ for every $v \in V$. A minimalistic version of this generalization would be to allow $c^t : V \to \{-1, 0, 1\}$, define the dynamic process by

$$c_v^{t+1} = sign\left(\sum_{w \in N(v)} c_w^t\right),$$

and state that G admits a majority consensus for c^0 if there exists a t such that c^t is a consensus with $(c_v^t) = sign(sum(c^0))$ for every $v \in V$. (Note that the local majority process differs from this generalization only in the tie-breaking rule: If initial colorings are restricted to $c^0 : V \to \{-1, 1\}$, then $c_v^1 = 0$ if and only if the number of 1's and -1's is equal in $N(v)$.) For both generalizations it is straightforward to generate results that show that G cannot be an m.c.c. if there exists a partition of G similar to that described in Theorem 3.2(b). For example, no bipartite graph can be an m.c.c. in either of the two generalizations.

7. Conclusions and directions.

The main result of this paper is that failure-free computation of majority consensus by iterative applications of the local majority rule is possible only in the networks that are nowhere truly local (Theorem 3.5). In other words, the idea of solving a truly global task (reaching consensus on majority) by means of truly local computation only (local majority rule) is doomed for failure.

However, even well connected networks of agents that are nowhere truly local might fail to reach majority consensus when iteratively applying the local majority rule. We have investigated the properties of m.c.c.'s, i.e., the networks in which iterative application of the local majority rule always yields consensus in the initial majority state.

There are several directions that might be of potential interest. One direction that was not of our interest involves computational issues, such as determining the computational complexity of the decision problem:

DMCC (Deciding an M.C.C.). Input is a finite graph G. Is G an m.c.c.?

Clearly, DMCC is in co-NP because of Theorem 3.2, and it is very likely that DMCC is co-NP complete. However, subclasses of DMCC are in P; cf. Corollary 4.14. Another possible direction could be extremal properties of m.c.c.'s. For example, it would be interesting to determine what is the minimal number of edges in an m.c.c. on n vertices. Our results provide only an obvious lower bound of n (Corollary 3.3) and an upper bound of $\binom{n}{2} - \binom{(n-1)/2}{2}$ edges (Proposition 3.1(a)).

The direction that would be more along the lines of our work would be a quest for the full characterization of m.c.c.'s. We have made a first step towards a possible characterization theorem by characterizing m.c.c.'s for networks that are almost complete in the sense that every agent does not communicate with at most two other agents (Theorem 4.12). A simpler task would be to determine interesting properties of m.c.c.'s that fall short of characterization. For example, we have shown, by an exhaustive computer aided search, that in every m.c.c. on at most 13 agents there exists an agent that communicates with all other agents. In fact, we conjecture that every m.c.c. G contains a master; i.e., there exists $v \in V(G)$ such that $d(v) = |V(G)| - 1$ (see the master conjecture in section 3). We have shown that this conjecture holds for almost complete networks, i.e., networks that are in a way natural candidates for a counterexample to the conjecture (Theorem 4.10). However, the master conjecture remains open.

REFERENCES

[1] H. Attiya and J. Welch, *Distributed Computing: Fundamentals, Simulations and Advanced Topics*, McGraw-Hill, New York, 1998.

[2] E. Berger, *Dynamic monopolies of constant size*, J. Combin. Theory Ser. B, 83 (2001), pp. 191–200.

[3] G. Bracha, *An $o(\log n)$ expected rounds randomized byzantine generals algorithm*, J. ACM, 34 (1987), pp. 910–920.

[4] S. B. Davidson, H. Garcia-Molina, and D. Skeen, *Consistency in partitioned networks*, ACM Computing Surveys, 17 (1985), pp. 341–370.

[5] M. H. Degroot, *Reaching a consensus*, J. Amer. Statist. Assoc., 69 (1974), pp. 167–182.

[6] K. Diks and D. Pelc, *System diagnosis with smallest risk of error*, Theoret. Comput. Sci., 203 (1998), pp. 163–173.

[7] P. Dreyer, *Applications and Variations of Domination in Graphs*, Ph.D. thesis, Rutgers University, New Brunswick, NJ, 2000.

[8] C. Dwork, D. Peleg, N. Pippenger, and E. Upfal, *Fault tolerance in networks of bounded degree*, SIAM J. Comput., 17 (1988), pp. 975–988.

[9] F. Flocchini, E. Lodi, F. Luccio, L. Pagli, and N. Santoro, *Monotone dynamos in tori*, in Proceedings of the 6th International Colloqium on Structural Information and Communication Complexity, Carleton Scientific, Ontario, Canada, 1999, pp. 152–165.

[10] P. Flocchini, E. Lodi, F. Luccio, and N. Santoro, *Irreversible dynamos in tori*, in Proceedings of the 4th International Euro-Par Conference on Parallel Processing, Lecture Notes in Comput. Sci., Springer-Verlag, Berlin, New York, 1998, pp. 554–562.

[11] J. French, *A formal theory of social power*, Psychology Rev., 63 (1956), pp. 181–194.

[12] H. Garcia-Molina and D. Barbara, *How to assign votes in a distributed system*, J. Assoc. Comput. Mach., 32 (1985), pp. 841–860.

[13] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman, San Francisco, CA, 1979.

[14] D. K. Gifford, *Weighted voting for replicated data*, in Proceedings of the 7th ACM Symposium on Operating Systems Principles, ACM Press, Pacific Grove, CA, 1979, pp. 150–162.

[15] E. Goles, *Positive automata networks*, in Disordered Systems and Biological Organization, Springer-Verlag, Berlin, New York, 1986, pp. 101–112.

[16] E. Goles and S. Martinez, *Neural and Automata Networks*, Kluwer, Norwell, MA, 1990.

[17] E. Goles and J. Olivos, *Periodic behavior of generalized threshold functions*, Discrete Math., 30 (1980), pp. 187–189.

[18] E. Goles and J. Olivos, *Comportement périodique des fonctions à seuil binaires et applications*, Discrete Appl. Math., 3 (1981), pp. 93–105.

[19] F. Harary, *A criterion for unanimity in French's theory of social power*, in Studies in Social Power, Inst. Soc. Res., University of Michigan Press, Ann Arbor, MI, 1959, pp. 168–182.

[20] Y. Hassin, *Probabilistic Local Polling Processes in Graphs*, M.Sc. thesis, The Weizmann Institute, Israel, 1998.

[21] Y. Hassin and D. Peleg, *Distributed probabilistic polling and applications to proportionate agreement*, Inform. and Comput., 171 (2001), pp. 248–268.

[22] Y. Hassin and D. Peleg, *Extremal bounds for probabilistic polling in graphs*, in Proceedings of the 7th International Colloqium on Structural Information and Communication Complexity, Carleton Scientific, Ontario, Canada, 2000, pp. 167–180.

[23] M. P. Herlihy, *Replication Methods for Abstract Data Types*, Ph.D. thesis, MIT, Cambridge, MA, 1984.

[24] S. Kutten and D. Peleg, *Tight fault locality*, SIAM J. Comput., 30 (2000), pp. 247–268.

[25] S. Kutten and D. Peleg, *Fault-local distributed mending*, J. Algorithms, 30 (1999), pp. 144–165.

[26] L. Lamport, R. Shostak, and M. Pease, *The byzantine generals problem*, ACM Trans. Programming Languages and Systems, 4 (1982), pp. 382–401.

[27] F. Luccio, L. Pagli, and H. Sanossian, *Irreversible dynamos in butterflies*, in Proceedings of the 6th International Colloqium on Structural Information and Communication Complexity, Carleton Scientific, Ontario, Canada, 1999, pp. 204–218.

[28] G. Moran, *Parametrization for stationary patterns of the r-majority operators on* 0–1 *sequences*, Discrete Math., 132 (1994), pp. 175–195.

[29] G. Moran, *The r-majority vote action on* 0–1 *sequences*, Discrete Math., 132 (1994), pp. 145–174.

[30] G. Moran, *On the period-two-property of the majority operator in infinite graphs*, Trans. Amer. Math. Soc., 347 (1995), pp. 1649–1667.

[31] N. H. Mustafa and A. Pekeč, *Majority consensus and the local majority rule*, in Proceedings ICALP '01, Lecture Notes in Comput. Sci. 2076, Springer-Verlag, Berlin, New York, 2001, pp. 530–542.

[32] T. Nakata, H. Imahayashi, and M. Yamashita, *Probabilistic local majority voting for the agreement problem on finite graphs*, in Proceedings of the 5th Computing and Combinatorics Conference, Lecture Notes in Comput. Sci. 1627, Springer, Berlin, 1999, pp. 330–338.

[33] T. Nakata, H. Imahayashi, and M. Yamashita, *A probabilistic local polling game on weighted directed graphs with an application to the distributed agreement problem*, Networks, 35 (2000), pp. 266–273.

[34] D. Pelc, *Efficient fault location with small risk*, in Proceedings of the 3rd International Colloquium on Structural Information and Communication Complexity, Carleton Scientific, Ontario, Canada, 1996, pp. 292–300.

[35] D. Peleg, *Local majority voting, small coalitions, and controlling monopolies in graphs: A review*, in Proceedings of the 3rd International Colloquium on Structural Information and Communication Complexity, Carleton Scientific, Ontario, Canada, 1996, pp. 152–169.

[36] D. Peleg, *Size bounds for dynamic monopolies*, Discrete Appl. Math., 86 (1998), pp. 263–273.

[37] D. Peleg and A. Wool, *The availability of quorom systems*, Inform. Comput., 12 (1995), pp. 210–223.

[38] S. Poljak and M. Sura, *On periodical behaviour in societies with symmetric influences*, Combinatorica, 3 (1983), pp. 119–121.

[39] S. Poljak and D. Turzik, *On an application of convexity to discrete systems*, Discrete Appl. Math., 13 (1986), pp. 27–32.

[40] S. Poljak and D. Turzik, *On pre-periods of discrete influence systems*, Discrete Appl. Math., 13 (1986), pp. 33–39.

[41] M. Spasojevic and P. Berman, *Geometric voting as the optimal pessimistic scheme for managing replicated data*, IEEE Trans. Parallel Distrib. Systems, 5 (1994), pp. 64–73.

[42] G. Sullivan, *The Complexity of System-Level Fault Diagnosis and Diagnosability*, Ph.D. thesis, Department of Computer Science, Yale University, New Haven, CT, 1986.

[43] A. Wool, *Quorom Systems for Distributed Control Protocols*, Ph.D. thesis, The Weizmann Institute, Israel, 1996.

SIAM J. Discrete Math.
Vol. 17, No. 4, pp. 661–669

FRACTIONAL PACKING OF T-JOINS*

FRANCISCO BARAHONA†

Abstract. Given a graph with nonnegative capacities on its edges, it is well known that the capacity of a minimum T-cut is equal to the value of a maximum fractional packing of T-joins. The Padberg–Rao algorithm finds a minimum capacity T-cut, but it does not produce a T-join packing. We present a polynomial combinatorial algorithm for finding an optimal T-join packing.

Key word. T-joins

AMS subject classification. 90C27

DOI. 10.1137/S0895480102409619

1. Introduction.

We present a polynomial combinatorial algorithm for packing T-joins in a capacitated graph. Given a graph $G = (V, E)$ and $S \subseteq V$, the set of all edges with exactly one endnode in S is called a *cut* and is denoted by $\delta_G(S)$. We say that S *defines* the cut $\delta_G(S)$. If the graph G is clear from the context, we use $\delta(S)$. Given a set $T \subseteq V$ of even cardinality, we say that a cut $\delta(S)$ is a T-cut if $|S \cap T|$ is odd. A set of edges J is called a *T-join* if, in the subgraph $G' = (V, J)$, the nodes in T have odd degree and the nodes in $V \setminus T$ have even degree. T-joins appear in the solution to the Chinese postman problem of Edmonds and Johnson [5]. Here the nodes in T are the nodes of odd degree, and a T-join is a set of edges that have to be duplicated to obtain an Eulerian graph.

Edmonds and Johnson [5] proved that if A is a matrix whose rows are the incidence vectors of all T-cuts, then for any nonnegative objective function w the linear program below has an optimal integer solution that is the incidence vector of a T-join.

$$\min wx, \tag{1}$$
$$Ax \geq 1, \tag{2}$$
$$x \geq 0. \tag{3}$$

Edmonds and Johnson gave a combinatorial polynomial algorithm to solve the linear program above and its dual

$$\max y1, \tag{4}$$
$$yA \leq w, \tag{5}$$
$$y \geq 0. \tag{6}$$

This gives a packing of T-cuts. Seymour [15] proved that if the coefficients of w are integer, and their sum over every cycle is an even number, then (4)–(6) have an optimal integer solution. The algorithm of Edmonds and Johnson can be modified to produce this integer dual optimal solution; see [2].

It follows from the theory of blocking polyhedra [6] that if B is a matrix whose rows are all incidence vectors of T-joins, then for any nonnegative objective function

*Received by the editors June 17, 2002; accepted for publication (in revised form) November 28, 2003; published electronically May 17, 2004.

http://www.siam.org/journals/sidma/17-4/40961.html

†IBM T. J. Watson Research Center, Yorktown Heights, NY 10589 (barahon@us.ibm.com).

c the linear program below also has an optimal integer solution that is the incidence vector of a T-cut:

$$\min cx, \tag{7}$$

$$Bx \geq 1, \tag{8}$$

$$x \geq 0. \tag{9}$$

The dual problem is

$$\max y1, \tag{10}$$

$$yB \leq c, \tag{11}$$

$$y \geq 0. \tag{12}$$

A solution of (10)–(12) is a maximum packing of T-joins. So from linear programming duality, we have that the value of a maximum packing of T-joins is equal to the value of a minimum T-cut. Padberg and Rao [13] gave a polynomial combinatorial algorithm that finds a minimum T-cut. However, this algorithm does not give a maximum packing of T-joins, which has remained unsolved. Due to the equivalence between separation and optimization, one could solve this in polynomial time with the ellipsoid method; see [10]. The purpose of this paper is to give a polynomial combinatorial algorithm for finding a maximum (fractional) packing. To the best of our knowledge, the only case that is well solved is when $|T| = 2$; this is the well-known maximum flow problem. Our algorithm has many similarities with an algorithm for packing arborescences given by Gabow and Manu [8].

There are several conjectures and questions related to the case when the linear program (10)–(12) has an integer solution. We discuss them below.

A graph is called *r-regular* if all its vertices have degree r. A graph is called an *r-graph* if it is r-regular and every V-cut has cardinality greater than or equal to r. A *perfect matching* is a set of nonadjacent edges that covers every vertex of the graph. Fulkerson made the following conjecture.

Conjecture 1. *Every 3-graph has 6 perfect matchings that include each edge at most twice.*

Notice that for a 3-graph, when $T = V$ every vertex defines a minimum T-cut. Also every T-join with positive weight in a maximum packing should intersect a minimum T-cut in exactly one edge, so the T-join should be a perfect matching. Thus in our terminology the conjecture above is equivalent to saying that for a 3-graph when $T = V$ and c is a vector of all 2's, then (10)–(12) has an optimal solution that is integer.

Seymour [14] generalized Fulkerson's conjecture as follows.

Conjecture 2. *Every r-graph has $2r$ perfect matchings that include each edge at most twice.*

Seymour [14] also made the following two conjectures and proved that they are implied by Conjecture 2. A family of T-joins is called *k-disjoint* if every edge is included in at most k of them.

Conjecture 3. *If every vertex has an even degree, then the size of a maximum 2-disjoint family of T-joins equals twice the size of a minimum T-cut.*

Conjecture 4. *The size of a 4-disjoint family of T-joins equals four times the size of a minimum T-cut.*

Cohen and Lucchesi [3] made the conjecture below and proved that it is equivalent to Conjecture 2.

CONJECTURE 5. *If all T-cuts have the same parity, then the size of a maximum 2-disjoint family of T-joins equals twice the size of a minimum T-cut.*

They also proved the following.

THEOREM 1. *If $|T| \leq 8$ and every T-cut has the same parity, then the size of a maximum disjoint family of T-joins equals the size of a minimum T-cut.*

Conforti and Johnson [4] made the following conjecture. They proved their conjecture for graphs without a 4-wheel minor.

CONJECTURE 6. *If T is the set of nodes of odd degree, and the graph is not contractible to the Petersen graph, then the size of a maximum disjoint family of T-joins equals the size of a minimum T-cut.*

Holyer [11] proved that deciding whether a 3-regular simple graph has 3 disjoint perfect matchings is NP-complete. So finding an optimal integer solution of (10)–(12) is NP-hard. Tait [16] proved that the 4-color theorem is equivalent to the statement that every 2-connected planar 3-regular graph has 3 disjoint perfect matchings. This is equivalent to saying that for every 2-connected planar 3-regular graph, when $T = V$ and c is the vector of all 1's, the linear program (10)–(12) has an optimal solution that is integer.

Now we give some extra notation and definitions. Let $n = |V|$ and $m = |E|$. We assume that every edge e has a nonnegative *capacity* $c(e)$. If $c(e)$ is zero, then the edge e is removed from the graph. For $S \subseteq V$ we use $\theta(S)$ to denote

$$\theta(S) = \sum\{c(e) \,:\, e \in \delta(S)\};$$

this is the *capacity* of the cut $\delta(S)$. Given $A, B \subseteq V$, we say that they *cross* if the sets $A \setminus B$, $B \setminus A$, and $A \cap B$ are nonempty. A family of sets such that no two of them cross is called *laminar*. A laminar family of subsets of V can have at most $2n - 1$ nonempty sets. It is well known that θ is a *submodular* function; i.e., for any two sets $A, B \subseteq V$,

$$\theta(A \cup B) + \theta(A \cap B) + 2\beta(A, B) = \theta(A) + \theta(B),$$

where $\beta(A, B)$ is the sum of the capacities of the edges with one endnode in $A \setminus B$ and the other in $B \setminus A$. We use $\lambda(G)$ to denote the capacity of a minimum T-cut in G, i.e.,

$$\lambda(G) = \min\{\theta(S) \,:\, S \subset V,\ |S \cap T| \text{ is odd}\}.$$

For $U \subseteq E$ we use $\mu(U)$ to denote

$$\mu(U) = \min\{c(e) \,:\, e \in U\};$$

this is called the *bottleneck capacity* of U. If J is a T-join and $\delta(S)$ is a cut, then $|J \cap \delta(S)|$ is odd if and only if $\delta(S)$ is a T-cut. If $U \subseteq E$ and $0 \leq \alpha \leq \mu(U)$, we denote by $G - \alpha U$ the graph obtained by replacing the capacity $c(e)$ of every edge $e \in U$ with $c(e) - \alpha$. If $U \subseteq E$ the *incidence vector* of U, denoted by x^U, is defined as $x^U(e) = 1$ if $e \in U$, and $x^U(e) = 0$ otherwise. A minimum cut separating nodes s and t is called a *minimum st-cut.* The nodes in the set T are called *T-nodes.*

This paper is organized as follows. In section 2 we give a short description of the Padberg–Rao algorithm for finding a minimum T-cut. In section 3 we present an initial description of the algorithm for packing T-joins. Sections 4 and 5 are devoted to more technical aspects required to complete the description of our algorithm. Section 6 contains a final analysis of our algorithm.

2. The Padberg–Rao algorithm. For the sake of completeness, we give a short description of the Padberg–Rao algorithm for finding a minimum T-cut. It is based on the following lemma.

LEMMA 1. *Let S define a minimum cut separating at least two nodes in T. If $|S \cap T|$ is odd, then S defines a minimum T-cut. Otherwise, there is a set $S' \subseteq S$ or $S' \subseteq V \setminus S$ that defines a minimum T-cut.*

Proof. Assume that $|S \cap T|$ is even and consider a set A that defines a minimum T-cut. Suppose that A and S cross.

Case 1. $|A \cap S \cap T|$ is odd. If $A \cup S$ separates at least two nodes in T, we have

$$\theta(A \cap S) + \theta(A \cup S) \leq \theta(A) + \theta(S).$$

Therefore $\theta(A \cap S) = \theta(A)$ and $\theta(A \cup S) = \theta(S)$. Thus $A \cap S$ defines a minimum T-cut.

If $T \subseteq A \cup S$, let $\bar{A} = V \setminus A$; then

$$\theta(\bar{A} \cap S) + \theta(\bar{A} \cup S) \leq \theta(\bar{A}) + \theta(S).$$

Thus $\theta(\bar{A} \cap S) = \theta(\bar{A})$, $\theta(\bar{A} \cup S) = \theta(S)$, and $\bar{A} \cap S$ defines a minimum T-cut.

Case 2. $|A \cap S \cap T|$ is even. Let $\bar{S} = V \setminus S$. Then $|A \cap \bar{S} \cap T|$ is odd and this reduces to Case 1. □

This lemma suggests a very simple algorithm, namely if S defines a minimum cut separating at least two nodes in T, then either S defines a minimum T-cut or one should continue working recursively with the graph G_1, obtained by contracting S, and with the graph G_2, obtained by contracting $V \setminus S$.

Padberg and Rao also pointed out that one should first compute a Gomory–Hu (GH) tree [9] and then carry out the algorithm above on the GH-tree. This is because any minimum st-cut in the graph is given by a minimum st-cut in the GH-tree. Because of the tree structure, the algorithm becomes extremely simple: among all edges in the tree that are T-cuts, we should choose one of minimum capacity.

Thus the complexity of this procedure is the complexity of computing a GH-tree, i.e., computing $(n-1)$ minimum st-cuts.

3. The algorithm. We start this section with an initial description of the algorithm. Clearly, the capacity of any T-cut is an upper bound for the value of a T-join packing. As mentioned in the introduction, there is a fractional packing of T-joins whose value is equal to the capacity of a minimum T-cut. For this bound to be tight, any T-join with a positive weight in an optimal packing must intersect any minimum T-cut in exactly one edge. Also, given an optimal packing, every edge e in a minimum T-cut, with $c(e) > 0$, must appear in a T-join with positive weight. The algorithm works based on this.

Using $\lambda(G)$ as the target value, the problem is solved recursively in a *greedy* way as follows. For a T-join U, let α_U be the largest value of α such that $\lambda(G - \alpha U) = \lambda(G) - \alpha$ and $0 \leq \alpha \leq \mu(U)$. Then the weight α_U is assigned to U. If $\lambda(G - \alpha_U U) > 0$, one should continue working recursively with $G - \alpha_U U$. In the remainder of this paper we show that a refinement of this algorithm runs in polynomial time. We need first a simple lemma.

LEMMA 2. *If U is a T-join and $\alpha_U = 0$, then there is a minimum T-cut $\delta(S)$ such that $|\delta(S) \cap U| > 1$.*

Proof. First notice that $\lambda(G - \alpha U) \leq \lambda(G) - \alpha$ for $0 \leq \alpha \leq \mu(U)$. This is because in $G - \alpha U$ the capacity of a T-cut $\delta(S)$ is $\theta(S) - k\alpha$, where $k = |\delta_G(S) \cap U|$.

So if $|\delta(S) \cap U| = 1$ for every minimum T-cut $\delta(S)$, then there is a small value of $\alpha > 0$ such that $\lambda(G - \alpha U) = \lambda(G) - \alpha$ and $\alpha \leq \mu(U)$. □

From the lemma above we can see that one should concentrate on T-joins that intersect *every* minimum T-cut in exactly *one* edge. When we impose this condition for a minimum T-cut $\delta(S)$, we say that it is *tight*; we also say that S is a *tight set.* The two lemmas below show that we only need to impose this for a laminar family of tight sets.

LEMMA 3. *Assume that A and B define minimum T-cuts, they cross, and $|A \cap B \cap T|$ is odd. Then the tightness of $A \cap B$ and $A \cup B$ implies the tightness of A and B.*

Proof. We have that

$$\theta(A \cap B) + \theta(A \cup B) \leq \theta(A) + \theta(B).$$

Since A and B define minimum T-cuts, then $A \cap B$ and $A \cup B$ also define minimum T-cuts. Therefore this inequality must hold as an equation. This implies that there is no edge between $A \setminus B$ and $B \setminus A$. Moreover, for a T-join U and any cut $\delta(S)$ the cardinality of $\delta(S) \cap U$ is odd if S defines a T-cut and even otherwise. Then by a counting argument it is easy to see that any T-join that has exactly one edge entering $A \cap B$ and exactly one edge entering $A \cup B$ must have exactly one edge entering A and exactly one edge entering B. Figure 1 displays all possible configurations. □

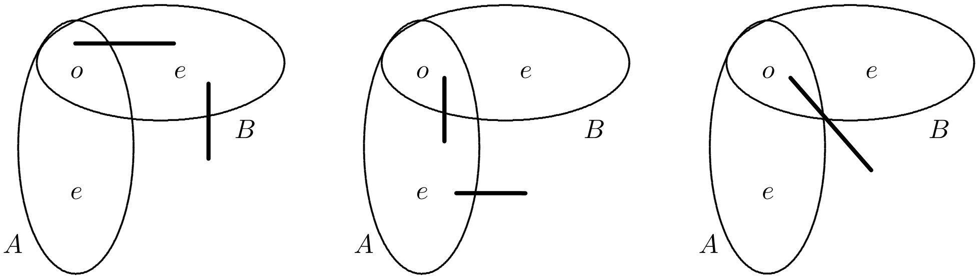


FIG. 1. *The labels e (even) and o (odd) refer to the parity of* $|(A \setminus B) \cap T|$, $|A \cap B \cap T|$, *and* $|(B \setminus A) \cap T|$.

LEMMA 4. *Assume that A and B define minimum T-cuts, they cross, and $|A \cap B \cap T|$ is even. Then the tightness of $A \setminus B$ and $B \setminus A$ implies the tightness of A and B.*

Proof. Apply Lemma 3 to A and $\bar{B} = V \setminus B$. □

So when we keep a family of tight sets, we can apply the last two lemmas to convert it into a laminar family. Denote this family by Φ; it can contain at most $2n - 1$ tight sets. We are going to find a T-join that intersects every T-cut given by Φ in exactly one edge. Let U be this T-join. There are two possible cases as follows:

1. If $\alpha_U = \mu(U)$, then the number of edges in $G - \alpha_U U$ is at least one less than the number of edges in G.

2. If $\alpha_U < \mu(U)$, then in $G - \alpha_U U$ there is a minimum T-cut $\delta(S)$, $S \notin \Phi$, such that $|U \cap \delta(S)| > 1$. In this case we should add S to Φ and uncross it using Lemmas 3 and 4 as in the procedure below.

Uncross (Φ, *S*, *U*)

```
Input: The family $\Phi$, a set $S \notin \Phi$, a $T$-join $U$.
While there is a set $A \in \Phi$ such that $A$ and $S$ cross
do
    if $|A \cap S \cap T|$ is odd
        if $|\delta(A \cup S) \cap U| > 1$, set $S \leftarrow A \cup S$
        if $|\delta(A \cup S) \cap U| = 1$, set $S \leftarrow A \cap S$
    if $|A \cap S \cap T|$ is even
        if $|\delta(A \setminus S) \cap U| > 1$, set $S \leftarrow A \setminus S$
        if $|\delta(A \setminus S) \cap U| = 1$, set $S \leftarrow S \setminus A$
end
Add $S$ to $\Phi$.
```

It is easy to see that at each uncrossing step the number of crossing pairs decreases by at least one. Also, at the end of this procedure the cardinality of Φ increases by one.

Now we can give a formal description of the algorithm.

Pack *T*-joins

```
Step 0. Set $\Phi \leftarrow \emptyset$.
Step 1. Find a $T$-join $U$ such that $|U \cap \delta(S)| = 1$ for all $S \in \Phi$.
Step 2. Compute $\alpha_U$ as the maximum of $\alpha$ such that
$\lambda(G - \alpha U) = \lambda(G) - \alpha$ and $0 \le \alpha \le \mu(U)$.
Step 3. If $\alpha_U < \mu(U)$, a new tight $T$-cut $\delta(S)$ has been found.
Apply Uncross($\Phi$, $S$, $U$).
Step 4. Set $G \leftarrow G - \alpha_U U$. If $\lambda(G) = 0$, stop; otherwise go to Step 1.
```

Since at each iteration either the cardinality of Φ increases or one edge is deleted, the total number of iterations is at most $2n - 1 + m$. It remains to describe how to perform Steps 1 and 2. This is the subject of the next two sections.

4. Finding a *T*-join in Step 1.

As it was said in the introduction, it follows from linear programming duality that there is a fractional packing of T-joins whose value is $\lambda(G)$. The purpose of this section is to find one T-join that is a candidate for having positive weight in the optimal packing. We start with some properties of these T-joins.

LEMMA 5. *Let $\delta(S)$ be a minimum T-cut; then every T-join with positive weight in an optimal packing intersects $\delta(S)$ in exactly one edge.* □

LEMMA 6. *Let $\delta(S)$ be a minimum T-cut and $e \in \delta(S)$ with $c(e) > 0$. Let $\{U_i\}$ be the set of T-joins in an optimal packing with weights $y(U_i) > 0$ for all i. Then there is at least one T-join U_i such that $U_i \cap \delta(S) = \{e\}$. Moreover,*

$$c(e) = \sum_{U_i \,:\, e \in U_i} y(U_i). \qquad \square$$

LEMMA 7. *Let $\delta(S)$ be a minimum T-cut. Let G' be the graph obtained by shrinking S to a single node and giving it the label T. Let G'' be the graph obtained by shrinking $V \setminus S$ to a single node and giving it the label T. An optimal packing of T-joins in G can be obtained by combining the elements of an optimal packing in G' with the elements of an optimal packing in G''.*

Proof of Lemma 7. Clearly $\lambda(G') = \lambda(G'') = \lambda(G)$. Let $\{U'_i\}$ be the family in an optimal packing in G' with weights $y'(U'_i) > 0$ for all i. Let $\{U''_j\}$ be the family in an optimal packing in G'' with weights $y''(U''_j) > 0$ for all j. Consider $e \in \delta(S)$ with $c(e) > 0$. We have

$$c(e) = \sum_{U'_i \,:\, e \in U'_i} y'(U'_i) = \sum_{U''_j \,:\, e \in U''_j} y''(U''_j).$$

Thus if U'_i and U''_j contain the edge e, then their union gives a T-join U. We give it the weight $y(U) = \min\{y'(U'_i), y''(U''_j)\}$; then the value $y(U)$ is subtracted from $y'(U'_i)$ and $y''(U''_j)$ and, if any of these weights becomes 0, the corresponding T-join is removed. This is repeated for any pair with positive weights containing the edge e.

Then this procedure is applied for every edge $e \in \delta(S)$. □

Given the family Φ of tight sets, we need to find a T-join U such that $|U \cap \delta(S)| = 1$ for all $S \in \Phi$. Lemma 7 suggests that the graph should be decomposed using minimum T-cuts, and it shows how to combine T-joins from the pieces. The procedure is described below.

First, we start with $S = V$ and define G_S as the subgraph induced by S, with every maximal set of Φ that is properly contained in S contracted, labeled as a T-node, and marked as *tight*. Let T_S be the set of T-nodes in G_S. We define an auxiliary graph whose node set is T_S; this is a complete graph. For any two nodes in T_S we find a path in G_S between them of minimum cardinality. Tight nodes can be the beginning or the end of a path, but cannot be intermediate nodes. This is to ensure that the resulting T-join intersects every tight T-cut exactly once. The cardinality of this path becomes the weight of the corresponding edge in the auxiliary graph. We give infinite weight if the path does not exist. We find a minimum weight perfect matching in the auxiliary graph. This is to ensure that the resulting T-join is minimal. In G_S we take the union of all paths whose corresponding edges are in the matching. This gives a T-join U' in G_S. Every tight node has exactly one edge of U' incident to it. Lemmas 5 and 6 show that a matching of finite weight exists; any T-join with positive weight in an optimal packing produces a matching of finite weight in the auxiliary graph.

Then we have to deal with each set W that has been contracted. In the T-join above, there is exactly one edge $e = \{i, j\}$ with $j \in W$. This time G_S is the subgraph induced by W plus the edge e, and the node i is labeled as a T-node. The idea is to find a T-join U'' in this new graph and combine U' and U'' as in Lemma 7. Again every maximal set of Φ that is properly contained in $S = W \cup \{i\}$ is contracted, and we proceed as above. The algorithm continues recursively. This produces a T-join U that is a candidate for appearing in an optimal packing; its weight α_U is obtained as in the next section.

The complexity of finding a minimum weight perfect matching in a complete graph with t nodes is $O(t^3)$; see [7, 12]. Also, the complexity of finding all shortest paths in G_S is $O(t^3)$. Therefore the complexity of Step 1 is $O(n^3)$.

5. Finding α_U in Step 2.

Given a T-join U, we compute the maximum value of α such that

$$\lambda(G - \alpha U) = \lambda(G) - \alpha \quad \text{and} \quad 0 \le \alpha \le \mu(U).$$

Let us define $f(\alpha) = \lambda(G - \alpha U)$. The function f is the minimum of a set of affine linear functions, and so it is concave and piecewise linear. We have to find its first breakpoint. For this we start with a tentative value $\alpha_U = \mu(U)$. We compute $f(\alpha_U)$. If $f(\alpha_U) = \lambda(G) - \alpha_U$, we are done; otherwise let $\delta(S)$ be a minimum T-cut in $G - \alpha_U U$. Let $k = |U \cap \delta_G(S)|$. Notice that here we use $\delta_G(S)$ because in

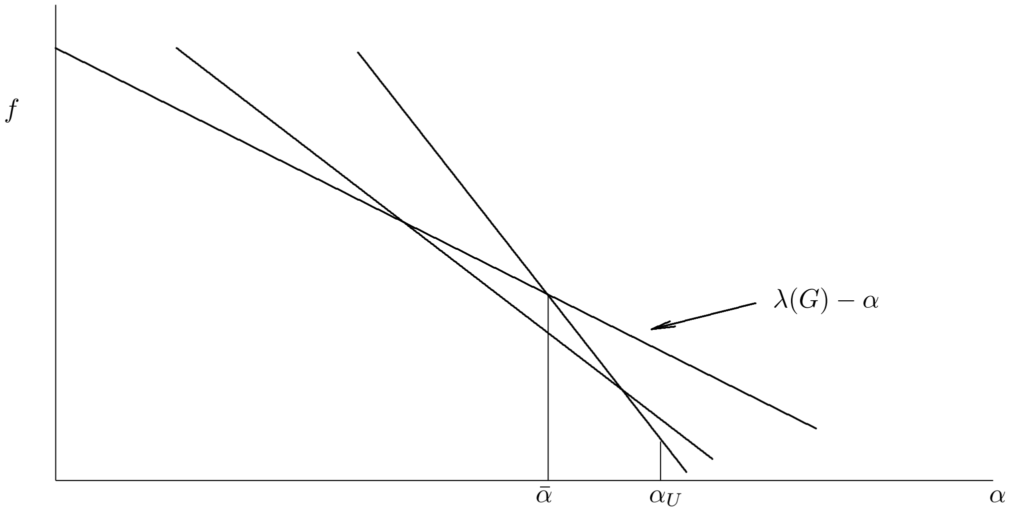


FIG. 2. *Finding α_U.*

$G - \alpha_U U$ we might have deleted some edges with capacity 0. Let $\bar{\alpha}$ be the solution of $\lambda(G) - \alpha = \theta(S) - k\alpha$. We set $\alpha_U \leftarrow \bar{\alpha}$ and continue. See Figure 2.

A formal description of this algorithm follows.

```
Find $\alpha_U$
Step 0. Set $\alpha_U \leftarrow \mu(U)$.
Step 1. Find a minimum $T$-cut $\delta(S)$ in $G - \alpha_U U$. If $\lambda(G - \alpha_U U) = \lambda(G) - \alpha_U$,
stop. Otherwise continue.
Step 2. Compute $\bar{\alpha}$ as the solution of $\lambda(G) - \alpha = \theta(S) - k\alpha$,
where $k = |U \cap \delta_G(S)|$.
Step 3. Set $\alpha_U \leftarrow \bar{\alpha}$ and go to Step 1.
```

The complexity of this algorithm is given below.

LEMMA 8. *If $\alpha_U = \mu(U)$, this algorithm requires $O(n)$ minimum st-cut computations; otherwise it requires $O(n^2)$ minimum st-cut computations.*

Proof. If $\alpha_U = \mu(U)$, only one iteration is performed. Otherwise at each iteration the value of $k = |U \cap \delta_G(S)|$ decreases. Since $|U| \leq n - 1$, the above algorithm takes at most $n - 1$ iterations. At each iteration one has to find a minimum T-cut with the Padberg–Rao algorithm; this requires $n - 1$ minimum st-cut computations. Then the result follows. □

6. Final analysis. Clearly, the running time of the algorithm in section 3 is dominated by the running time of Steps 1 and 2. Also notice that at most $2n - 1 + m$ iterations are performed. Thus the total running time of Step 1 is $O((n + m)n^3)$. For Step 2 there are at most m iterations, where we have $\alpha_U = \mu(U)$ that require $n - 1$ minimum st-cuts and have at most $2n - 1$ iterations that require at most $(n - 1)^2$ minimum st-cuts. The complexity of finding a minimum st-cut is $O(n^3)$; see [1]. Thus the total running time of Step 2 is $O((mn + n^3)n^3)$. Therefore the complexity of this algorithm is $O(n^6)$.

Since at each iteration one new T-join is produced, we have the following.

THEOREM 2. *There exists an optimal packing with at most $2n - 1 + m$ T-joins having a positive weight.* □

A vector $\bar{x}$ satisfying (2) and (3) can be decomposed into $\bar{x} = g + h$, where g is a convex combination of incidence vectors of T-joins and h is a nonnegative vector. This convex combination can be obtained as follows. Use the values $\bar{x}(e)$ as capacities, find an optimal packing of T-joins. Let $\{U_i\}$ be the family of T-joins with weights $y(U_i) > 0$. Let $\alpha = \sum y(U_i)$. Set $y'(U_i) = y(U_i)/\alpha$ for all i; then the vector g is

$$g = \sum y'(U_i) x^{U_i}.$$

Acknowledgment. I am grateful to an anonymous referee for several helpful comments.

REFERENCES

[1] R. K. AHUJA, T. L. MAGNANTI, AND J. B. ORLIN, *Network Flows. Theory, Algorithms, and Applications*, Prentice–Hall, Englewood Cliffs, NJ, 1993.

[2] F. BARAHONA, *Planar multicommodity flows, max cut, and the Chinese postman problem*, in Polyhedral Combinatorics, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 1, AMS, Providence, RI, 1990, pp. 189–202.

[3] J. COHEN AND C. LUCCHESI, *Minimax relations for T-join packing problems*, in Proceedings of the Fifth Israeli Symposium on Theory of Computing and Systems (ISTCS '97), 1997, pp. 38–44.

[4] M. CONFORTI AND E. JOHNSON, *Two Min-Max Theorems for Graphs Not Contractible to a 4-Wheel*, Technical report, IBM T. J. Watson Research Center, Yorktown Heights, NY, 1987.

[5] J. EDMONDS AND E. L. JOHNSON, *Matching, Euler tours and the Chinese postman*, Math. Programming, 5 (1973), pp. 88–124.

[6] D. R. FULKERSON, *Blocking and anti-blocking pairs of polyhedra*, Math. Programming, 1 (1971), pp. 168–194.

[7] H. N. GABOW, *An efficient implementation of Edmonds' algorithm for maximum matching on graphs*, J. Assoc. Comput. Mach., 23 (1976), pp. 221–234.

[8] H. N. GABOW AND K. S. MANU, *Packing algorithms for arborescences (and spanning trees) in capacitated graphs*, Math. Programming Ser. B, 82 (1998), pp. 83–109.

[9] R. E. GOMORY AND T. C. HU, *Multi-terminal network flows*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 551–570.

[10] M. GRÖTSCHEL, L. LOVÁSZ, AND A. SCHRIJVER, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, 1993.

[11] I. HOLYER, *The NP-completeness of edge-coloring*, SIAM J. Comput., 10 (1981), pp. 718–720.

[12] E. L. LAWLER, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, Montreal, 1976.

[13] M. W. PADBERG AND M. R. RAO, *Odd minimum cut-sets and b-matchings*, Math. Oper. Res., 7 (1982), pp. 67–80.

[14] P. D. SEYMOUR, *On multicolourings of cubic graphs, and conjectures of Fulkerson and Tutte*, Proc. London Math. Soc. (3), 38 (1979), pp. 423–460.

[15] P. D. SEYMOUR, *On odd cuts and plane multicommodity flows*, Proc. London Math. Soc. (3), 42 (1981), pp. 178–192.

[16] P. G. TAIT, *On the colouring of maps*, Proc. Roy. Soc. Edinburgh, 10 (1880), pp. 501–503.

SIAM J. DISCRETE MATH.
Vol. 17, No. 4, pp. 670–674

EXACT FORMULAE FOR THE LOVÁSZ THETA FUNCTION OF SPARSE CIRCULANT GRAPHS*

VALENTINO CRESPI[†]

Abstract. The Lovász theta function has attracted much attention for its connection with diverse issues such as communicating without errors and computing large cliques in graphs. Indeed, this function enjoys the remarkable property of being computable in polynomial time despite being *sandwiched* between clique and chromatic numbers, two well-known, hard to compute quantities.

In this paper I provide a closed formula for the Lovász function of all the circulant graphs of degree 4 with even displacement, thus generalizing Lovász results on cycle graphs (circulant graphs of degree 2).

Key words. Lovász theta function, circulant graph, linear programming

AMS subject classifications. 05C10, 68R10, 90C05, 90C22

DOI. 10.1137/S089548010241852X

1. Introduction.

Consider a graph G whose vertices represent letters from a given alphabet, and where adjacency indicates that two letters can be "confused." The zero-error *capacity* of G is the number $\Theta(G)$ of messages that can be communicated without any error. This notion was introduced in 1956 by Shannon [6] and has generated much interest over the years. It was understood quite early that the exact determination of the Shannon capacity is a very difficult problem, even for small and simple graphs. In 1979 Lovász [5] introduced a related function, which soon thereafter became known as the Lovász theta function, or Lovász number, with the aim of estimating the Shannon capacity.

The Lovász theta function (denoted by $\vartheta(G)$) is computable in polynomial time, although it is "sandwiched" between the clique number $\omega(G)$ and the chromatic number $k(G)$, whose computation is NP-hard. Because of this remarkable property, and also because of its relevance to communication issues, the Lovász number is widely studied (see the survey by Knuth [4]).

Despite much work in the field, very little is known about classes of graphs whose theta function can be expressed with formulae involving a constant "low" number of simple operations (e.g., arithmetic, logarithmic, and/or trigonometric). A rare example of such a result is Lovász's formula for n-cycles with n odd [5]:

$$\vartheta(C_n) = \frac{n\cos\frac{\pi}{n}}{1+\cos\frac{\pi}{n}}.$$

Recently Brimkov et al. [2, 3] obtained formulae for the more general cases of circulant graphs with chord lengths 2 and 3.

In this paper I use a geometric approach to establish and prove a closed formula for the theta function of circulant graphs of degree 4 when the displacement j is even. The formula itself was already suggested in [3] but was not fully proved.

*Received by the editors November 22, 2002; accepted for publication (in revised form) October 22, 2003; published electronically May 17, 2004.

http://www.siam.org/journals/sidma/17-4/41852.html

[†]Department of Computer Science, College of Engineering, Computer Science and Technology, California State University, Los Angeles, CA 90032 (vcrespi@calstatela.edu). Part of this work was done when the author was in service at Dartmouth College, Hanover, NH.

Here I close the issue establishing that, for j even and $n > 2(j+1)j$, the following holds:

$$\vartheta(C_{n,j}) = \frac{n}{1 + \frac{\cos\left(\frac{2\pi(i+1)j}{n}\right) - \cos\left(\frac{2\pi(i+1)}{n}\right) + \cos\left(\frac{2\pi i}{n}\right) - \cos\left(\frac{2\pi ij}{n}\right)}{\cos\left(\frac{2\pi(i+1)}{n}\right)\cdot\cos\left(\frac{2\pi ij}{n}\right) - \cos\left(\frac{2\pi i}{n}\right)\cdot\cos\left(\frac{2\pi(i+1)j}{n}\right)}} \tag{1}$$

with $i = \lfloor \frac{nj}{2(j+1)} \rfloor$.

This paper, together with the results appearing in [3] for the complementary case j odd, provides a full analysis of the theta function of all the circulant graphs of degree 4.

2. Preliminaries.

2.1. Some graph-theoretical notions and facts. Let us recall some well-known definitions from graph theory. Given a graph $G(V, E)$, its *complement graph* is the graph $\bar{G}(V, \bar{E})$, where $\bar{E}$ is the complement of E to the set of edges of the complete graph on V. An *automorphism* of the graph G is a permutation p of its vertices such that two vertices $u, v \in V$ are adjacent if and only if $p(u)$ and $p(v)$ are adjacent. G is *vertex symmetric* if its automorphism group is vertex transitive; i.e., for any given $u, v \in V$ there is an automorphism p such that $p(u) = v$.

A graph $G' = (V', E')$ is an *induced subgraph* of $G(V, E)$ if E' contains all edges from E that join vertices from $V' \subseteq V$. G is called *perfect* if $\omega(G_{V'}) = k(G_{V'})$ for all $V' \subseteq V$, where $G_{V'}$ is the induced subgraph of G on the vertex set V'.

An $n \times n$ matrix $A = (a_{i,j})_{i,j=0}^{n-1}$ is called *circulant* if its entries satisfy $a_{i,j} = a_{0,j-i}$, where the subscripts belong to the set $\{0, 1, \ldots, n-1\}$ and are calculated modulo n. In other words, any row of a circulant matrix can be obtained from the first one by a number of consecutive cyclic shifts, and thus the matrix is fully determined by its first row. A *circulant graph* is a graph with a circulant adjacency matrix. The expression $C_{n,j}$ will denote a circulant graph of degree 4, with vertex set $\{0, 1, \ldots, n-1\}$ and edge set $\{(i, i+1 \bmod n), (i, i+j \bmod n), i = 0, 1, \ldots, n-1\}$, where $1 < j \leq \lfloor \frac{n}{2} \rfloor$ is the *chord length.*

Several equivalent definitions of the Lovász number are available [4]. Presented here is one which requires only little technical machinery.

DEFINITION 2.1. *Given a graph G, let $\mathbf{A}$ be the family of matrices A such that $a_{ij} = 0$ if v_i and v_j are adjacent in G. Let $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$ be the eigenvalues of A. Then $\vartheta(G) = \max_{A \in \mathbf{A}}\{1 - \frac{\lambda_1(A)}{\lambda_n(A)}\}$.*

An important property to recall is the following.

PROPOSITION 2.2 (see [5]). *For every graph G with n vertices, $\vartheta(G) \cdot \vartheta(\bar{G}) \geq n$. If G is vertex symmetric, then $\vartheta(G) \cdot \vartheta(\bar{G}) = n$.*

2.2. Linear programming (LP) formulation. Taking advantage of the particular properties of circulant matrices, whose eigenvalues can be expressed in closed formulae, and so generalizing the approach in [5], we can easily derive the validity of the following minmax formulation of the theta function of circulant graphs of degree 4.

LEMMA 2.3 (see [2]). *Let $f_0(x, y) = n + 2x + 2y$ and, for some fixed value of j, $f_i(x, y) = 2x \cos \frac{2\pi i}{n} + 2y \cos \frac{2\pi ij}{n}$, $i = 1, 2, \ldots, n-1$. Then*

$$\vartheta(C_{n,j}) = \min_{x,y} \max_{i} \left\{ f_i(x, y), i = 0, 1, \ldots, \left\lfloor \frac{n}{2} \right\rfloor \right\}. \tag{2}$$

This in turn is equivalent to the following LP problem:

$$\vartheta(C_{n,j}) = \min\left\{z \ : \ f_i(x,y) - z \le 0,\ i = 0, 1, \ldots, \left\lfloor \frac{n}{2} \right\rfloor, z \ge 0\right\}. \tag{3}$$

Through a nontrivial analysis of the structure of the admissible region defined by the linear constraints, it was possible for us to obtain closed formulae for some special cases of circulant graphs of degree 4 [2]. For example, I report the least complex formula for the simplest case $j = 2$:

$$\vartheta(C_{n,2}) = n\left(1 - \frac{\frac{1}{2} - \cos(\frac{2\pi}{n}\lfloor\frac{n}{3}\rfloor) - \cos(\frac{2\pi}{n}(\lfloor\frac{n}{3}\rfloor + 1))}{(\cos(\frac{2\pi}{n}\lfloor\frac{n}{3}\rfloor) - 1)(\cos(\frac{2\pi}{n}(\lfloor\frac{n}{3}\rfloor + 1)) - 1)}\right). \tag{4}$$

3. Proof of the formula for $\boldsymbol{\vartheta(C_{n,j})}$. The general idea is to identify the indices of the constraints in LP problem (3) that determine the optimal vertex as functions of n and j only. The result was originally obtained through a direct study of the geometric regularities of the admissible region of the primal problem [1]. This led to a rather complex and long proof. A simpler and more concise way to derive the formula, as suggested by one of the referees, is to analyze the dual of the LP problem (3) as explained below.

THEOREM 3.1. *Let n and j be integer numbers. Assume that j is even and $n > 2(1 + j)j$. Then $\vartheta(C_{n,j}) = z_0$, where (x_0, y_0, z_0) is the only solution to the following 3×3 linear system:*

$$\begin{cases} n + 2x + 2y & = & z, \\ 2x\cos(\frac{2\pi k}{n}) + 2y\cos(\frac{2\pi kj}{n}) & = & z, \\ 2x\cos(\frac{2\pi(k+1)}{n}) + 2y\cos(\frac{2\pi(k+1)j}{n}) & = & z \end{cases}$$

for $k = \lfloor\frac{nj}{2(j+1)}\rfloor$. By Cramer's rule this gives formula (1).

Proof. The dual of the LP problem (3) is

maximize nu_0
subject to

$$u_0, u_1, \ldots, n_{\lfloor n/2\rfloor} \ \ge\ 0, \tag{5}$$

$$\sum_{i=0}^{\lfloor n/2\rfloor} u_i \le 1, \tag{6}$$

$$\sum_{i=0}^{\lfloor n/2\rfloor} u_i \cdot 2\cos\frac{2\pi i}{n} \ = \ 0, \tag{7}$$

$$\sum_{i=0}^{\lfloor n/2\rfloor} u_i \cdot 2\cos\frac{2\pi ij}{n} \ = \ 0. \tag{8}$$

Constraint (6) can be replaced with $\sum_{i=0}^{\lfloor n/2\rfloor} u_i = 1$ because any admissible solution $(u'_0, u'_1, \ldots, u'_{\lfloor n/2\rfloor})$ for which $\sum_i u'_i < 1$ could be improved by setting $u''_i = u'_i/\sum_k u'_k$. Constraints (7) and (8) can be rewritten in vector notation as

$$\sum_{i=1}^{\lfloor n/2\rfloor} u_i \begin{bmatrix} \cos\frac{2\pi i}{n} \\ \cos\frac{2\pi ij}{n} \end{bmatrix} = -u_0 \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

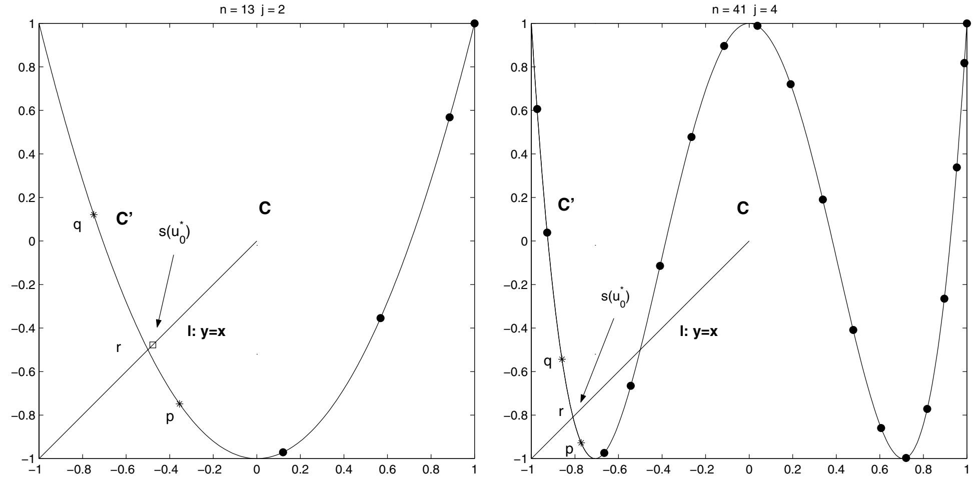


FIG. 1. *Function $f(x) = \cos(j \arccos x)$, $j = 2, 4$. The segment joining p and q identifies the solution $s(u_0^*)$ on the bisectrix line.*

and, after substituting $u_0 = 1 - \sum_{k=1}^{\lfloor n/2 \rfloor} u_k$, the result becomes

$$\sum_{i=1}^{\lfloor n/2 \rfloor} \frac{u_i}{\sum_{k=1}^{\lfloor n/2 \rfloor} u_k} \begin{bmatrix} \cos \frac{2\pi i}{n} \\ \cos \frac{2\pi ij}{n} \end{bmatrix} = -\frac{u_0}{1-u_0} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

This equation says that point $s(u_0) = (-\frac{u_0}{1-u_0}, -\frac{u_0}{1-u_0})$ is a convex combination of the points

$$X := \left\{ \left(\cos \frac{2\pi i}{n}, \cos \frac{2\pi ij}{n} \right) : i = 1, \dots, \left\lfloor \frac{n}{2} \right\rfloor \right\} .$$

So, if we denote by $CH(X)$ the convex hull of X, then $\vartheta(C_{n,j}) = nu_0^*$, where

$$u_0^* = \max_{0 \le u_0 \le 1} \{u_0 \mid s(u_0) \in CH(X)\} .$$

To determine this value, consider the curve

$$C := \{(\cos(\alpha), \cos(j\alpha)) \mid 0 \le \alpha \le \pi\}$$

and its subcurve

$$C' := \left\{ (\cos(\alpha), \cos(j\alpha)) \;\middle|\; \frac{j-1}{j}\pi \le \alpha \le \pi \right\} .$$

Curve C is the graph of the function $f : [-1, 1] \to [-1, 1]$ defined by $f(x) := \cos(j \arccos x)$. This function is convex on $[-1, \cos(\frac{j-1}{j}\pi)]$ and has its minimum value in $x_0 = \cos(\frac{j-1}{j}\pi)$. This implies that, for any two points p and q on C', the line through p and q separates the segment of C' connecting p and q from the rest of C (see Figure 1). Point $s(u_0)$ lies on line $l : y = x$ and, as we increase u_0 from 0 to 1, $s(u_0)$ slides from the origin $(0, 0)$ to $(-\infty, -\infty)$. Line l intersects curve C' in point $r = (\cos \alpha, \cos j\alpha)$ with $\alpha = \frac{j}{j+1}\pi$. This can be seen by solving the equation

$$\cos(j\alpha) - \cos(\alpha) = -2 \sin \frac{\alpha(j+1)}{2} \cdot \sin \frac{\alpha(j-1)}{2} = 0$$

for $\alpha \in [\frac{j-1}{j}\pi, \pi]$. So, setting $k = \lfloor \frac{nj}{2(j+1)} \rfloor$, points $p := (\cos\frac{2\pi k}{n}, \cos\frac{2\pi kj}{n})$ and $q := (\cos\frac{2\pi(k+1)}{n}, \cos\frac{2\pi(k+1)j}{n})$ are the points in X closest (along C) to r (in both directions). Since $n > 2(1+j)j$, both p and q belong to C'. The segment of C connecting p and q contains no other points in X, so $s(u_0^*)$ must be a convex combination of p and q. This implies that the optimal solution $U^* = (u_i^*)_{0 \le i \le \lfloor n/2 \rfloor}$ verifies $u_0^*, u_k^*, u_{k+1}^* > 0$ and $u_i^* = 0$ for $i \neq 0$, $i \neq k$, and $i \neq k+1$, and so by the complementary slackness theorem the three corresponding inequalities in the primal problem are tight, giving the 3×3 linear system in the theorem statement. □

Notes. The condition $n > 2(1+j)j$, given in Theorem 3.1 for the validity of the formula is only sufficient but not necessary. This means that, for some j, it can be weakened.

Acknowledgments. The author wishes to thank Bruno Codenotti for inspiring this work, and Valentin Brimkov and Luis Caffarelli for their important suggestions that have made it possible. The author also wishes to thank the anonymous referee who suggested the proof of Theorem 3.1 based on the dual of LP problem (3).

REFERENCES

[1] V. Crespi, *Exact Formulae for the Lovász Theta Function of Sparse Circulant Graphs*, Technical report TR2002-438, Computer Science Department, Dartmouth College, Hanover, NH, 2002.

[2] V. E. Brimkov, B. Codenotti, V. Crespi, and M. Leoncini, *On the Lovász number of certain circulant graphs*, in Algorithms and Complexity, Lecture Notes in Comput. Sci. 1767, G. Bongiovanni, G. Gambosi, and R. Petreschi, eds., Springer-Verlag, Berlin, New York, 2000, pp. 291–305.

[3] V. E. Brimkov, B. Codenotti, V. Crespi, R. P. Barneva, and M. Leoncini, *Efficient Computation of the Lovász Theta Function for a Class of Circulant Graphs*, manuscript.

[4] D. E. Knuth, *The sandwich theorem*, Electron. J. Combin., 1 (1994), pp. 1–48.

[5] L. Lovász, *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory, 25 (1979), pp. 1–7.

[6] C. E. Shannon, *The zero-error capacity of a noisy channel*, Institute of Radio Engineers Transactions on Information Theory, IT-2 (1956), pp. 8–19.

SIAM J. Discrete Math.
Vol. 17, No. 4, pp. 675–685

COUNTING STRINGS WITH GIVEN ELEMENTARY SYMMETRIC FUNCTION EVALUATIONS I: STRINGS OVER $\mathbb{Z}_p$ WITH p PRIME[*]

C. R. MIERS[†] AND F. RUSKEY[‡]

Abstract. Let α be a string over $\mathbb{Z}_p$ with p prime. The jth elementary symmetric function evaluated at α is denoted $T_j(\alpha)$. We study the cardinalities $S_p(n; \tau_1, \tau_2, \dots, \tau_t)$ of the set of length n strings for which $T_i(\alpha) = \tau_i$. The *profile* $\langle k_0, k_1, \dots, k_{p-1}\rangle$ of a string α is the sequence of frequencies with which each letter occurs. The profile of α determines $T_j(\alpha)$, and hence S_p. Let $f_n : \mathbb{Z}_{p^n}^{p-1} \mapsto \mathbb{Z}_p^{p^n-1}$ be the map that takes $\langle k_0, k_1, \dots, k_{p-1}\rangle \bmod p^n$ to $(T_1, T_2, \dots, T_{p^n-1}) \bmod p$. We show that f_n is well defined and injective and show how to efficiently determine its range. These results are used to efficiently compute $S_p(n; \tau_1, \tau_2, \dots, \tau_t)$.

Key words. elementary symmetric function, combinatorial enumeration, integers mod a prime

AMS subject classifications. 05A15, 05E05, 05A19

DOI. 10.1137/S0895480102415381

1. Introduction.

The theory of symmetric functions has long been a basic tool of combinatorial enumeration. Indeed, Cameron [1] states that "one can appreciate the view held by some people, that if it isn't related to symmetric polynomials, then it isn't combinatorics!". However, the enumeration of the number of variable assignments to symmetric functions so that the functions achieve given values seems to be new and interesting.

In order for the problem to make sense, we must choose the variables to come from some finite algebraic structure and choose a particular class of symmetric functions. Here we choose the variables to come from the ring of integers mod a prime p and choose the class of elementary symmetric functions. The elementary symmetric functions are important because they give the coefficients of polynomials in terms of their roots.

The main purpose of this paper, and its companion paper [6], is to count certain strings over the ring of integers mod p^n and over the finite field $\mathbb{F}_{p^n}$, where p is prime. We take the point of view espoused by Wilf [7] that the intrinsic worth of an expression is determined by the amount of computation that it takes to evaluate it. We will state our running times in terms of the number of ring and arithmetic operations that it takes to evaluate the expression using the obvious algorithm. The word size of the computer is assumed to be $O(\log n)$ since the largest numbers we deal with have size $O(r^n)$ where r, the cardinality of the ring, is regarded as a constant.

This work is a continuation of [2], where the number of monic irreducible polynomials over $\mathbb{F}_2$ of degree n with given trace and "subtrace" are enumerated. The trace is the coefficient of x^{n-1} and the subtrace the coefficient of x^{n-2}. If such a polynomial is factored in a splitting field, the trace and subtrace can be viewed as the first and second elementary symmetric functions evaluated at the string of coefficients

[*] Received by the editors September 27, 2002; accepted for publication (in revised form) December 8, 2003; published electronically May 17, 2004.
http://www.siam.org/journals/sidma/17-4/41538.html

[†] Department of Mathematics and Statistics, University of Victoria, Victoria, BC, V8W 3P6, Canada (crmiers@math.uvic.ca). This author's research was supported in part by a UVic Faculty Research grant.

[‡] Department of Computer Science, University of Victoria, Victoria, BC, V8W 3P6, Canada (fruskey@cs.uvic.ca). This author's research was supported in part by the NSERC.

appearing in the factorization. The techniques in [2] are elementary in nature and involve the relationship of Lyndon words to irreducible polynomials. It therefore seems a natural extension of these ideas to count higher order "traces" on strings with values in various rings.

2. Preliminaries.

Consider a string $\alpha = a_1 a_2 \cdots a_n$, where each $a_i \in \mathbb{Z}_p$. Define the *j-trace* of α, $T_j(\alpha)$, to be the sum

$$T_j(\alpha) = \sum_{1 \le i_1 < i_2 < \cdots < i_j \le n} a_{i_1} a_{i_2} \cdots a_{i_j} \pmod{p}.$$

These are the elementary symmetric functions evaluated at $a_1, a_2, \ldots, a_n$. Clearly, $(-1)^j T_j(\alpha)$ is the negation of the coefficient of z^{n-j} in the polynomial

$$(z - a_1)(z - a_2) \cdots (z - a_n).$$

By $S_p(n; \tau_1, \tau_2, \ldots, \tau_j)$ we denote the number of strings α over R of length n for which $T_i(\alpha) = \tau_i$ for $i = 1, 2, \ldots, j$. Obviously if $j = 0$, then $S_p(n) = r^n$. It is also true that $S_p(n; t) = p^{n-1}$ for any $t \in R$ since $T_1(\alpha x)$ takes on distinct values for each $x \in R$.

In what follows, the notation $[\![P]\!]$ for proposition P has the value 1 if P is true and the value 0 if P is false. This is "Iverson's convention" as used in [4].

The numbers $S_p(n; \tau_1, \tau_2, \ldots, \tau_t)$ satisfy the following recurrence relation. If $n = 1$, then $S_p(n; \tau_1, \tau_2, \ldots, \tau_j) = [\![\tau_2 = \cdots = \tau_j = 0]\!]$, and for $n > 0$,

$$S_p(n; \tau_1, \tau_2, \ldots, \tau_j) = \sum_{x \in \mathbb{Z}_p} S_p(n-1; \rho_1, \rho_2, \ldots, \rho_j), \tag{2.1}$$

where $\rho_0 = 1$, and for $i = 1, 2, \ldots, j$,

$$\rho_i = \tau_i - \rho_{i-1} x.$$

Iterating yields (with $\tau_0 = 1$)

$$\rho_i = \sum_{\ell=0}^{i} (-1)^\ell \tau_{i-\ell} x^\ell.$$

Recurrence relation (2.1) implies that the power series $\sum_{n \ge 0} S_p(n; \tau_1, \tau_2, \ldots, \tau_j) z^n$ is rational. We can evaluate $S_p(n; \tau_1, \tau_2, \ldots, \tau_j)$ by creating a table of size np^j consisting of S_p for all strings of length at most n and over all j-traces. Each table entry requires $\Theta(pj)$ ring operations and $\Theta(p)$ arithmetic operations for a total of $\Theta(njp^{j+1})$ ring operations and $\Theta(np^{j+1})$ arithmetic operations. An aim of this paper is to reduce the number of ring and arithmetic operations required to evaluate S_p. We begin in the next subsection by classifying the strings according to the frequency with which particular characters occur.

2.1. Profiles.

Suppose that the string α has k_x occurrences of the symbol x for $x \in \mathbb{Z}_p$. We refer to the $(p-1)$-tuple of natural numbers $\mathbf{k} = \langle k_1, k_2, \ldots, k_{p-1} \rangle$ as the *profile* of the string. Note that k_0 is omitted since it doesn't affect T_j. Subsequently, a bold letter will only denote a profile. We add profiles componentwise and define $r\mathbf{k} = \langle rk_1, rk_2, \ldots, rk_n \rangle$ for $r \in \mathbb{Z}_p$.

The j-trace T_j depends only on the profile, and we have

$$T_j(\alpha) \;=\; \sum_{\substack{\nu_1+\nu_2+\cdots+\nu_{r-1}=j\\ 0\le\nu_i\le k_i}} \prod_{i=1}^{r-1} i^{\nu_i}\binom{k_i}{\nu_i} \pmod p. \tag{2.2}$$

For $\mathbf{k}=\langle k_1,k_2,\ldots,k_{p-1}\rangle\in\mathbb{Z}^{p-1}$, define in $\mathbb{Z}_p[[z]]$ the formal power series

$$A_{\mathbf{k}}(z)=\prod_{j=1}^{p-1}(1+jz)^{k_j}. \tag{2.3}$$

We make no assumption here that the k_i's are positive.

Observe that

$$T_j(\alpha)=[z^j]A_{\mathbf{k}}(z), \tag{2.4}$$

where the notation $[z^j]A(z)$ indicates the coefficient of z^j in the generating function $A(z)$.

LEMMA 2.1.

$$A_{\mathbf{a}+\mathbf{b}}(z)=A_{\mathbf{a}}(z)A_{\mathbf{b}}(z). \tag{2.5}$$

Proof. The proof is clear. □

Throughout the rest of the paper, we assume that p is prime and set $\mathbb{Z}_p=\mathbb{Z}/p\mathbb{Z}$ and $\mathbb{Z}_{p^n}=\mathbb{Z}/p^n\mathbb{Z}$. We note that the characteristic of both of these rings is p.

THEOREM 2.2. *For all $n>0$,*

$$A_{p^n\mathbf{k}}(z)=A_{\mathbf{k}}(z^{p^n}).$$

Proof. Since p is prime and arithmetic is mod p, we have $(1+jz)^{p^n}=1+j^{p^n}z^{p^n}=1+jz^{p^n}$. Thus,

$$A_{p^n\mathbf{k}}(z)=\prod_{j=1}^{p-1}(1+jz)^{p^nk_j}=\prod_{j=1}^{p-1}(1+jz^{p^n})^{k_j}=A_{\mathbf{k}}(z^{p^n}). \qquad \square$$

COROLLARY 2.3. *For all $n>0$,*

$$A_{\mathbf{a}+p^n\mathbf{b}}(z)=A_{\mathbf{a}}(z) \bmod z^{p^n}.$$

Proof. The proof follows from Lemma 2.1, since $A_{\mathbf{a}+p^n\mathbf{b}}(z)=A_{\mathbf{a}}(z)A_{p^n\mathbf{b}}(z)=A_{\mathbf{a}}(z)A_{\mathbf{b}}(z^{p^n})=A_{\mathbf{a}}(z) \pmod{z^{p^n}}$. □

Notice that this corollary implies that, if we are considering only traces T_j with $j<p^n$, then we need only consider values of the profiles taken mod p^n.

We also denote the sum in (2.2) by $T_j(\mathbf{k})$ or $T_j(\langle k_1,k_2,\ldots,k_{p-1}\rangle)$ when we wish to emphasize the role of profiles. Let α and β be strings over $\mathbb{Z}_p$. The j-trace satisfies a natural convolution.

$$T_j(\alpha\beta) \;=\; \sum_{0\le i\le j} T_i(\alpha)T_{j-i}(\beta). \tag{2.6}$$

In terms of profiles, this becomes

$$T_j(\mathbf{k}+\mathbf{k}') \;=\; \sum_{0\le i\le j} T_i(\mathbf{k})T_{j-i}(\mathbf{k}'). \tag{2.7}$$

The evaluation of S_p in terms of profiles is given below:

$$(2.8)\quad S_p(n;\tau_1,\tau_2,\ldots,\tau_t) = \sum_{\substack{k_0+k_1+\cdots+k_{p-1}=n\\ \mathbf{k}:=\langle k_1,\ldots,k_{p-1}\rangle}} \binom{n}{k_0,k_1,\ldots,k_{p-1}} \prod_{i=1}^{t} [\![T_i(\mathbf{k})=\tau_i)]\!].$$

In order to evaluate (2.8) efficiently, we need to be able to determine efficiently those profiles $\mathbf{k}$ for which $T_i(\mathbf{k}) = \rho_i$ for $i = 1, 2, \ldots, t$. We will do this in the sections to follow.

3. The rings $\mathbb{Z}_p$ and $\mathbb{Z}_{p^n}$.

3.1. The fundamental correspondence. We first show that the map f that sends the $(p-1)$-tuple $\mathbf{k} = \langle k_1, k_2, \ldots, k_{p-1}\rangle$ to $\langle \tau_1, \tau_2, \ldots, \tau_{p-1}\rangle$, where $\tau_j = T_j(\mathbf{k}) = \sum \prod i^{\nu_i}\binom{k_i}{\nu_i}$, is a bijection on $\mathbb{Z}_p^{p-1}$.

LEMMA 3.1. *Let p be a prime and V be the $(p-1)\times(p-1)$ Vandermonde matrix defined by $v_{i,j} = j^i$ (mod p). Then $V^{-1} = W$ is the $(p-1)\times(p-1)$ matrix defined by $w_{i,j} = -i^{-j}$ (mod p).*

Proof. Let $c_{i,j}$ be the i, j entry of the matrix product VW:

$$c_{i,j} = -\sum_{1\le k<p} k^{i-j} = \begin{cases} 0 & \text{if } i \ne j,\\ -(p-1) & \text{if } i = j. \end{cases}$$

Thus $c_{i,j} = [\![i = j \bmod p]\!]$. The second equality follows from the proof of the first theorem about characters on finite Abelian groups as applied to the map $\chi : x \mapsto x^{i-j}$ on $\mathbb{Z}_p{}^* = \mathbb{Z}_p \setminus \{0\}$. That is, $\sum_{g\in G}\chi(g) = |G|\cdot[\![\chi \text{ is trivial}]\!]$ (see, e.g., [5, Theorem 5.4]). $\square$

Clearly f is a function; we will prove that it has an inverse f^{-1}. We refer to the result of the following theorem as the "fundamental correspondence."

THEOREM 3.2. *The map $f : \mathbb{Z}_p^{p-1} \mapsto \mathbb{Z}_p^{p-1}$ defined by $f(\mathbf{k}) = \langle\tau_1, \tau_2, \ldots, \tau_{p-1}\rangle$, where $\tau_i = T_i(\mathbf{k})$, is a bijection. Both f and f^{-1} can be computed in $O(p^2)$ arithmetic operations.*

Proof. The jth power symmetric function in variables $x_1, x_2, \ldots, x_t$, denoted $P_j(x_1, \ldots, x_t)$, is defined as

$$P_j(x_1, x_2, \ldots, x_t) = \sum_{i=1}^{t} x_i^j.$$

The Newton–Girard formula,

$$(3.1)\quad mT_m(x_1, x_2, \ldots, x_t) + \sum_{1\le j\le m} (-1)^j P_j(x_1, x_2, \ldots, x_t) T_{m-j}(x_1, x_2, \ldots, x_t) = 0,$$

allows us to express a power symmetric function as a (unique) polynomial of elementary symmetric functions. Given fixed values of the variables, $P_m = P_m(x_1, x_2, \ldots, x_t)$ and $T_m = T_m(x_1, x_2, \ldots, x_t)$ are values, and we can use (3.1) to compute unique values $P_1, P_2, \ldots, P_r$ from $T_1, T_2, \ldots, T_r$ by iterating the following equation for $m = 1, 2, \ldots, r$ (in that order). The successive computation of $P_1, P_2, \ldots, P_r$ will clearly take a total of $\Theta(p^2)$ arithmetic steps:

$$P_m = (-1)^{m+1}\left(mT_m + \sum_{1\le j\le m-1} (-1)^j P_j T_{m-j}\right).$$

Note that, as a function of the profile, $P_j = P_j(\langle k_1, k_2, \ldots, k_{p-1}\rangle) = \sum_{i=1}^{p-1} k_i i^j$. We therefore have the system of linear equations $\langle P_1, P_2, \ldots, P_{p-1}\rangle^T = V_p\langle k_1, k_2, \ldots, k_{p-1}\rangle^T$, where V_p is the $(p-1)\times(p-1)$ Vandermonde matrix with $V_p[i,j] = j^i \bmod p$. Since the Vandermonde matrix is nonsingular, this system has a unique solution $\langle k_1, k_2, \ldots, k_{p-1}\rangle$, thereby showing that f^{-1} is a function, as claimed. Further, the explicit expression for V_p^{-1} given in Lemma 3.1 allows us to compute the profile in $\Theta(p^2)$ arithmetic operations in $\mathbb{Z}_p$. □

The corollary below follows at once from Theorem 3.2 and the observation that $T_i(0, 0, \ldots, 0) = 0$ for any $i > 0$.

COROLLARY 3.3. *If* $T_i(\mathbf{k} \bmod p) = 0$ *for* $i = 1, 2, \ldots, p-1$*, then* $k_1 = k_2 = \cdots = k_{p-1} = 0$.

Example 1. Let us determine, in $\mathbb{Z}_7$, the profile that corresponds to the trace values $(T_1, T_2, \ldots, T_{p-1}) = (1, 1, 1, 1, 1, 1)$. The Newton–Girard formula can be written as

$$\begin{bmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ P_1 & -2 & & & & \\ P_2 & -P_1 & 3 & & & \\ P_3 & -P_2 & P_1 & -4 & & \\ P_4 & -P_3 & P_2 & -P_1 & 5 & \\ P_5 & -P_4 & P_3 & -P_2 & P_1 & -6 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Solving by back substitution, we get $(P_1, P_2, P_3, P_4, P_5, P_6)^T = (1, 6, 1, 6, 1, 6)^T$.

We now solve $(1, 6, 1, 6, 1, 6)^T = V_7\langle k_1, k_2, \ldots, k_{p-1}\rangle$,

$$\begin{bmatrix} 1 \\ 6 \\ 1 \\ 6 \\ 1 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 4 & 2 & 2 & 4 & 1 \\ 1 & 1 & 6 & 1 & 6 & 6 \\ 1 & 2 & 4 & 4 & 2 & 1 \\ 1 & 4 & 5 & 2 & 3 & 6 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ k_3 \\ k_4 \\ k_5 \\ k_6 \end{bmatrix},$$

by using the inverse V_7^{-1} computed from Lemma 3.1,

$$\begin{bmatrix} k_1 \\ k_2 \\ k_3 \\ k_4 \\ k_5 \\ k_6 \end{bmatrix} = \begin{bmatrix} 6 & 6 & 6 & 6 & 6 & 6 \\ 3 & 5 & 6 & 3 & 5 & 6 \\ 2 & 3 & 1 & 5 & 4 & 6 \\ 5 & 3 & 6 & 5 & 3 & 6 \\ 4 & 5 & 1 & 3 & 2 & 6 \\ 1 & 6 & 1 & 6 & 1 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 6 \\ 1 \\ 6 \\ 1 \\ 6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 6 \end{bmatrix}.$$

Thus the number of strings α of length n over $\mathbb{Z}_7$ with $T_j(\alpha) = 1$ for $j = 1, 2, 3, 4, 5, 6$ is

$$S_{\mathbb{Z}_7}(n; 1, 1, 1, 1, 1, 1) = \sum_{\substack{k_0 + k_1 + \cdots + k_6 = n \\ k_1 \equiv \cdots \equiv k_5 \equiv 0 \wedge k_6 \equiv 6 \pmod 7}} \binom{n}{k_0, k_1, \ldots, k_6}. \tag{3.2}$$

The actual values for $n = 1, 2, \ldots, 20$ are 0, 0, 0, 0, 0, 1, 7, 28, 84, 210, 462, 924, 10,297, 123,137, 906,010, 4,813,368, 20,435,156, 73,540,572, 232,846,824, 1,996,062,481. This computation takes a couple of seconds in Maple, after rearranging (3.2) into the form

$$\sum_{m=0}^{\lfloor (n-6)/7 \rfloor} \binom{n}{7m+6} \sum_{\substack{\nu_1+\cdots+\nu_6=m \\ \nu_i \geq 0}} \binom{7m+6}{7\nu_1,\ldots,7\nu_5,7\nu_6+6}.$$

Note that the number of terms in the above sum is about $\binom{(n/7)+6}{7}$.

Using classical results about primitive roots of unity (see the appendix) we can express (3.2) as a sum of 7^6 terms, each term of which raises a complex number to the power n. Equation (3.2) can be written as

$$\frac{1}{7^6} \sum_{\nu_1=0}^{6} \cdots \sum_{\nu_6=0}^{6} \omega^{\nu_6} (1+\omega^{\nu_1}+\cdots+\omega^{\nu_6})^n, \tag{3.3}$$

where ω is a primitive 7th root of unity. In infinite precision complex arithmetic, we can evaluate sums such as (3.3) in time $\Theta(p^{p-1}\log n)$ by using binary powering. However, in Maple the computation using (3.3) is much slower for realistic values of n.

3.2. Extending the fundamental correspondence. In this subsection we will prove that the map $f_n : \mathbb{Z}_{p^n}^{p-1} \mapsto \mathbb{Z}_p^{p^n-1}$ that sends $\mathbf{k} = \langle k_1, k_2, \ldots k_{p-1}\rangle \bmod p^n$ to $\langle \tau_1, \tau_2, \ldots, \tau_{p^n-1}\rangle \bmod p$, where $\tau_j = T_j(\mathbf{k})$, is one-to-one and determine its range for all $n \geq 2$. Let $\mathcal{P}_j = \{p^j, 2p^j, \ldots, (p-1)p^j\}$. We call the union $\mathcal{R}_m = \mathcal{P}_0 \cup \mathcal{P}_1 \cup \cdots \cup \mathcal{P}_m$ the *critical set* for the sequence $T_1, T_2, \ldots, T_{p^m-1}$; the elements of $\mathcal{R}_m$ are called *critical indices*. In extending the fundamental correspondence, we will prove that the map f_m, restricted to the values T_j where $j \in \mathcal{R}_m$, is a bijection. The values of T_j where j is not critical are determined by the values of T_i on the critical indices $i < j$. In the previous subsection we showed that f_1 is a bijection on $\mathbb{Z}_p^{p-1}$:

$$\underbrace{\underbrace{\underbrace{\overbrace{1,2,\ldots,p-1}^{\mathcal{P}_1\text{, taken mod }p}}_{f_1},\overbrace{p,\ldots,2p,\ldots,(p-1)p}^{\mathcal{P}_2\text{, taken mod }p^2}}_{f_2},\ldots,\overbrace{p^{m-1},\ldots,2p^{m-1},\ldots,(p-1)p^{m-1},\ldots,}^{\mathcal{P}_m\text{, taken mod }p^m}}_{f_m}p^m.$$

Lemma 3.4. *$A_{\mathbf{a}}(z) = A_{\mathbf{b}}(z) \bmod z^{p^n}$ if and only if $\mathbf{a} \equiv \mathbf{b} \bmod p^n$.*

Proof. If $\mathbf{a} \equiv \mathbf{b} \bmod p^n$, then by Corollary 2.3, $A_{\mathbf{a}}(z) = A_{\mathbf{b}}(z) \pmod{z^{p^n}}$.

Conversely, assume that $A_{\mathbf{a}}(z) = A_{\mathbf{b}}(z) \bmod z^{p^n}$. Then by (2.5), $A_{\mathbf{a}-\mathbf{b}}(z) = 1 \bmod z^{p^n}$. We proceed by induction on n. If $n = 1$, then by the fundamental correspondence, $\mathbf{a} \equiv \mathbf{b} \bmod p$. If $n > 1$, then we may assume inductively that $\mathbf{a} \equiv \mathbf{b} \bmod p^{n-1}$. Thus there is some $\mathbf{k} \in \mathbb{Z}_p^{p-1}$ such that $\mathbf{a} = \mathbf{b} + p^{n-1}\mathbf{k}$, and thus

$$1 = A_{\mathbf{a}-\mathbf{b}}(z) = A_{p^{n-1}\mathbf{k}}(z) = A_{\mathbf{k}}(z^{p^{n-1}}) \pmod{z^{p^n}}.$$

Since the condition $1 = A_{\mathbf{k}}(z^{p^{n-1}}) \bmod z^{p^n}$ is equivalent to $1 = A_{\mathbf{k}}(z) \bmod z^p$, by applying the fundamental correspondence, we obtain $\mathbf{k} = \mathbf{0} \bmod p$. □

Theorem 3.5. *The function f_n is one-to-one.*

Proof. We now assume that $\mathbf{k} \in \mathbb{Z}_{p^n}^{p-1}$. Lemma 3.4 shows that f_n (i.e., $A_{\mathbf{k}}(z) \bmod z^{p^n}$ regarded as a function of $\mathbf{k}$) is an injection. □

3.3. The range of f_n.

THEOREM 3.6. *The range of* $f_n : \mathbb{Z}_{p^n}^{p-1} \mapsto \mathbb{Z}_p^{p^n-1}$ *consists of all vectors*

$$\langle a_1, \dots, a_{p^{n-1}-1}, a_{p^{n-1}}, \dots, a_{2p^{n-1}}, \dots, a_{(p-1)p^{n-1}}, \dots, a_{p^n-1} \rangle,$$

where

(i) *the vector* $\langle a_1, \dots, a_{p^{n-1}-1} \rangle \in \mathrm{Range}(f_{n-1})$.
(ii) *the values of* $a_{mp^{n-1}}$ *can be assigned arbitrarily from* $\mathbb{Z}_p$ *for* $m = 1, 2, \dots, p-1$.
(iii) *for such an assignment, there are unique vectors* $\mathbf{a} \in \mathbb{Z}_p^{p-1}$ *and* $\mathbf{b} \in \mathbb{Z}_{p^{n-1}}^{p-1}$ *such that* $T_j(\mathbf{b}) = a_j$ *for* $j = 1, 2, \dots, p^{n-1} - 1$ *and* $T_{mp^{n-1}}(p^{n-1}\mathbf{a} + \mathbf{b}) = a_{mp^{n-1}}$ *for* $m = 1, 2, \dots, p-1$.
(iv) *the* a_j *for* $(m-1)p^{n-1} < j < mp^{n-1}$, $1 < m \le p$, *are determined uniquely as* $a_j = T_j(p^{n-1}\mathbf{a} + \mathbf{b})$.

Proof. Our proof is by induction on n. Given $(a_1, \dots, a_{p^{n-1}-1}) \in \mathrm{Range}(f_{n-1})$, there is a unique vector $\mathbf{b} \in \mathbb{Z}_{p^{n-1}}^{p-1}$ such that $[z^\ell]A_{\mathbf{b}}(z) = T_\ell(\mathbf{b}) = a_\ell$ for $\ell = 1, 2, \dots, p^{n-1} - 1$. Write $\mathbf{k} = p^{n-1}\mathbf{a} + \mathbf{b}$. If $1 \le \ell < p^{n-1}$, then $[z^\ell]A_{\mathbf{k}}(z) = [z^\ell](A_{p^{n-1}\mathbf{a}+\mathbf{b}}(z) \bmod p^{n-1}) = [z^\ell]A_{\mathbf{b}}(z)$. We also have

$$\begin{aligned}[z^{mp^{n-1}}]A_{\mathbf{k}}(z) &= [z^{mp^{n-1}}]A_{p^{n-1}\mathbf{a}+\mathbf{b}}(z) \\ &= [z^{mp^{n-1}}]A_{\mathbf{a}}(z^{p^{n-1}})A_{\mathbf{b}}(z) \\ &= \sum_{0 \le j \le m} T_{p^{n-1}j}(\mathbf{b})T_{m-j}(\mathbf{a}).\end{aligned}$$

Thus we are led to consider the equations

$$a_{mp^{n-1}} = \sum_{0 \le j \le m} T_{p^{n-1}j}(\mathbf{b})x_{m-j},$$

where $x_j = T_j(\mathbf{a})$. With $x_0 = 1$, we can uniquely determine the values $x_1, x_2, \dots, x_{p-1}$ successively by substitution in

$$x_m = a_{mp^{n-1}} - \sum_{1 \le j \le m} T_{p^{n-1}j}(\mathbf{b})x_{m-j}.$$

By the fundamental correspondence, the equations $x_j = T_j(\mathbf{a})$ for $j = 1, 2, \dots, p-1$ have a unique solution $\mathbf{a}$. Thus $\mathbf{k} = p^{n-1}\mathbf{a} + \mathbf{b}$ is a profile for which $a_i = T_i(\mathbf{k})$ for all $i \in \mathcal{R}_n = \mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \cup \mathcal{P}_n$. There are exactly $p^{n(p-1)}$ profiles of the form $p^{n-1}\mathbf{a} + \mathbf{b}$ and exactly $p^{n(p-1)}$ tuples a_i for $i \in \mathcal{R}_n$. Therefore, f_n is a bijection when restricted to $\mathcal{R}_n$. Furthermore, the values a_j for $j \in \{1, 2, \dots, p^n - 1\} \setminus \mathcal{R}_n$ are uniquely determined as $a_j = T_j(p^{n-1}\mathbf{a} + \mathbf{b})$. □

We showed above that the trace values are determined by the values of traces whose indices are in the critical set. We refine this below by showing that the value of T_t for t noncritical depends only on the values of T_j, where $j < t$ and j is critical.

THEOREM 3.7. *The value of* $T_t(\alpha)$, *where* $mp^{n-1} < t < (m+1)p^{n-1}$, *is determined by the values of* $\tau_j = T_j(\alpha)$ *for* $j \in \mathcal{R}_{n-1} \cup \{p^{n-1}, 2p^{n-1}, \dots, mp^{n-1}\}$.

Proof. By our previous results on the range of f_n, we know that there are exactly p^{p-1-m} profiles $\mathbf{k}$ such that $T_j(\mathbf{k}) = \tau_j$ for $j \in \mathcal{R}_{n-1} \cup \{p^{n-1}, 2p^{n-1}, \dots, mp^{n-1}\}$. Such a profile $\mathbf{k}$ can be written as $p^{n-1}\mathbf{a} + \mathbf{b}$, where $\mathbf{a} \in \mathbb{Z}_p^{p-1}$ and $\mathbf{b} \in \mathbb{Z}_{p^{n-1}}^{p-1}$.

Consider profiles $\mathbf{k} = p^{n-1}\mathbf{a} + \mathbf{b}$ and $\mathbf{k}' = p^{n-1}\mathbf{a}' + \mathbf{b}'$, where $T_j = \tau_j$ for $j \in \mathcal{R}_{n-1} \cup \{p^{n-1}, 2p^{n-1}, \ldots, mp^{n-1}\}$. Since the profiles agree on $\mathcal{R}_{n-1}$, we have $A_{\mathbf{k}}(z) = A_{\mathbf{k}'}(z) \bmod z^{p^{n-1}}$, and hence $\mathbf{b} = \mathbf{b}'$. Since, for $j = 1, 2, \ldots, m$,

$$[z^{jp^{n-1}}]A_{\mathbf{k}}(z) = [z^{jp^{n-1}}]A_{\mathbf{k}'}(z),$$

we have, for $j = 1, 2, \ldots, m$,

$$[z^{jp^{n-1}}]A_{\mathbf{a}}(z^{p^{n-1}})A_{\mathbf{b}}(z) = [z^{jp^{n-1}}]A_{\mathbf{a}'}(z^{p^{n-1}})A_{\mathbf{b}}(z),$$

which implies that

$$A_{\mathbf{a}}(z) = A_{\mathbf{a}'}(z) \pmod{z^{m+1}}.$$

Now note that

$$\begin{aligned}[z^t]A_{\mathbf{k}}(z) - [z^t]A_{\mathbf{k}}(z) &= [z^t](A_{\mathbf{a}}(z^{p^{n-1}}) - A_{\mathbf{a}'}(z^{p^{n-1}}))A_{\mathbf{b}}(z)\\ &= [z^t](A_{\mathbf{a}}(z^{p^{n-1}}) - A_{\mathbf{a}'}(z^{p^{n-1}}))A_{\mathbf{b}}(z) \pmod{z^{(m+1)p^{n-1}}}\\ &= 0. \qquad \square\end{aligned}$$

3.4. A computational method and examples. In this subsection we give an explicit algorithm in the form of pseudocode to determine if $\tau_1, \tau_2, \ldots, \tau_{p^n-1} \in \text{Range}(f_n)$, and, if so, how to find the profile $\mathbf{p} \in \mathbb{Z}_{p^n}^{p-1}$ such that $T_j(\mathbf{p}) = \tau_j$ for $j = 1, 2, \ldots, p^n - 1$. In particular, we will determine a sequence $\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_{n-1}$, where each $\mathbf{a}_i \in \mathbb{Z}_p^{p-1}$, such that

$$\mathbf{p} = \mathbf{a}_0 + p\mathbf{a}_1 + \cdots + p^{n-1}\mathbf{a}_{n-1}.$$

The principles underlying the algorithm have already been laid out in Theorems 3.2 and 3.6.

ALGORITHM.

```
(A1)   $\mathbf{a} := \mathbf{0}$;   $x_0 := 1$;
(A2)   for $i := 0$ to $n-1$ do
(A3)       for $j := 1$ to $p-1$ do
(A4)           $x_j := \tau_{jp^i} - \sum_{i=1}^{j} T_{jp^i}(\mathbf{a})x_{j-i}$;
(A5)       for $m := 1$ to $p-1$ do  { Newton–Girard }
(A6)           $P_m := (-1)^{m+1}\left(mx_m + \sum_{1\le j\le m-1}(-1)^j P_j x_{m-j}\right)$;
(A7)       for $j := 1$ to $p-1$ do  { inverse of Vandermonde }
(A8)           $a_j := \sum_{1\le i\le p-1}(-1)^{j+1}(p-i)^{p-j-1}P_i$;
(A9)       $\mathbf{a}_i := \mathbf{a}$;
(A10)  $\mathbf{p} := \mathbf{a}_0 + p\mathbf{a}_1 + \cdots + p^{n-1}\mathbf{a}_{n-1}$;
(A11)  for $i$ := 1 to $p^n - 1$ do
(A12)      if $i \notin \mathcal{R}_n$ and $T_i(\mathbf{p}) \neq \tau_i$ then return( "no profile exists" );
(A13)  return( $\mathbf{p}$ );
```

Example 2. Let us determine, in $\mathbb{Z}_{p^2}$, the profile $\mathbf{p}$, if any, that corresponds to the trace values $(T_1, T_2, \ldots, T_{p^2-1}) = (1, 1, \ldots, 1)$, with $p = 7$.

The $i = 0$ iteration of the algorithm was done in Example 1; $\mathbf{a} = \mathbf{a}_0 = (0, 0, 0, 0, 0, 6)$. For $i = 1$, the repeated substitution of lines (A3)–(A4) yields (with $\mathbf{b} = \mathbf{a}_1$)

$$(T_1(\mathbf{b}), T_2(\mathbf{b}), T_3(\mathbf{b}), T_4(\mathbf{b}), T_5(\mathbf{b}), T_6(\mathbf{b})) = (x_1, x_2, x_3, x_4, x_5, x_6) = (1, 1, 1, 1, 1, 1),$$

which is solved (lines (A5)–(A8)) as in the previous example to give $(b_1, b_2, b_3, b_4, b_5, b_6) = (0, 0, 0, 0, 0, 6)$. Thus $\mathbf{p} = 7\mathbf{b} + \mathbf{a} = (0, 0, 0, 0, 0, 48)$, where $48 = 7 \cdot b_6 + a_6 = 7{\cdot}6 + 6$. We now need to check, at lines (A11)–(A12), whether $T_j(\mathbf{p}) = 1$ for $7(m-1) < j < 7m$ for $m = 2, 3, 4, 5, 6, 7$. Consider a string of 48 6's. Clearly,

$$T_j(\mathbf{p}) = 6^j \binom{48}{j} = (-1)^j(-1)^j = 1 \pmod 7,$$

as long as $j \leq 48$. (To see that $\binom{48}{j} \equiv (-1)^j$, argue by induction using the recurrence relation $0 \equiv \binom{7^2}{j} = \binom{48}{j} + \binom{48}{j-1}$.) In terms of generating functions we have $(1-z)^{48} = 1 + z + \cdots + z^{48} \bmod 7$. Thus the T_j values are indeed all 1, and we can therefore determine that the number of strings of length n whose first 48 traces are all 1's is

$$S_{\mathbb{Z}_7}(n; \underbrace{1, 1, \ldots, 1}_{48}) = \sum_{\substack{k_0+k_1+\cdots+k_6=n \\ k_1\equiv\cdots\equiv k_5\equiv 0 \wedge k_6 \equiv 48 \pmod{7^2}}} \binom{n}{k_0, k_1, \ldots, k_6}. \tag{3.4}$$

Note that $7^{48} > 3 \times 10^{40}$, so there is no hope of using the recurrence relation (2.1) for the computation.

Example 3. Going in the other direction, specifying fewer traces to be 1, we show how to determine $S_{\mathbb{Z}_7}(n; 1, 1, 1)$. Since we don't have the complete set $\mathcal{P}_1$, we do not have a one-to-one correspondence. However, we can sum $S_{\mathbb{Z}_7}(n; 1, 1, 1, x, y, z)$ over all $x, y, z \in \mathbb{Z}_7$ to determine our answer. Feeding $(1, 1, 1, x, y, z)$ through the Newton–Girard and Vandermonde formulae gives us

$$\begin{aligned}
k_1 &\equiv 3 + 2x + 3y + 6z \pmod 7,\\
k_2 &\equiv 5 + 5x + 5y + 6z \pmod 7,\\
k_3 &\equiv 6 + 2x + 6z \pmod 7,\\
k_4 &\equiv 6 + 2y + 6z \pmod 7,\\
k_5 &\equiv 5 + 6x + 4y + 6z \pmod 7,\\
k_6 &\equiv 2 + 6x + 6y + 6z \pmod 7.
\end{aligned}$$

These equations can in turn be used to eliminate x, y, z, obtaining

$$\begin{aligned}
k_4 &\equiv k_1 + 4k2 + 3k3 \pmod 7,\\
k_5 &\equiv 3k_1 + 6k_2 + 6k_3 \pmod 7,\\
k_6 &\equiv 6 + 6k_1 + 6k_2 + 3k3 \pmod 7.
\end{aligned}$$

This gives us the equation

$$S_{\mathbb{Z}_7}(n; 1, 1, 1) = \sum_{\substack{k_0+k_1+\cdots+k_6=n \\ k_4\equiv k_1+4k2+3k3 \pmod 7 \\ k_5\equiv 3k_1+6k_2+6k_3 \pmod 7 \\ k_6\equiv 6+6k_1+6k_2+3k_3 \pmod 7}} \binom{n}{k_0, k_1, \ldots, k_6}.$$

Example 4. As another example, we will determine a formula for the number of binary strings of length n whose first 2^m traces are all 0's; i.e., we will determine the number

$$A(n, m) := S_{\mathbb{Z}_2}(n; \underbrace{0, 0, \ldots, 0}_{2^m}).$$

According to Theorem 3.6, the relevant trace values are T_j for $j = 1, 2, \ldots, 2^m$. From (A.1) of the appendix,

$$A(n,m) = \sum_{\substack{j\geq 0\\ j\equiv 0 \pmod{2^m}}} \binom{n}{j} = \frac{1}{2^m}\sum_{j=0}^{2^m-1}(1+\omega^j)^n = \frac{1}{2^m}\sum_{j=0}^{2^m-1}\left(2\cos\frac{\pi j}{2^m}\right)^n \cos\frac{\pi j n}{2^m},$$

where ω is a primitive 2^mth root of unity. The last equality follows from the observation that $(1+\omega^j) = \omega^{j/2}(\omega^{-j/2}+\omega^{j/2})$.

Appendix.

A.1. Roots of unity. For fixed n, it is well known that the sum of every other binomial coefficient is 2^{n-1}. But what about sums of every kth binomial coefficient? What about similar sums of multinomial coefficients? It turns out that we can derive formulae for these, whose computation is more efficient than directly summing the coefficients. We start with binomial coefficients and then proceed to the multinomial coefficients.

Let ω be a primitive qth root of unity, say $\omega = e^{2\pi i/q}$. Consider the geometric sum below for $q \nmid n$:

$$\sum_{k=0}^{q-1}\omega^{nk} = \frac{1-\omega^{qn}}{1-\omega^n} = 0.$$

On the other hand if $q \mid n$, then $\omega^{nk} = 1$. Thus

$$\frac{1}{q}\sum_{k=0}^{q-1}\omega^{nk} = [\![q \mid n]\!].$$

Let $A(z) = \sum_{n\geq 0} f(n)z^n$. We wish to find an expression for the related generating function that picks off every qth element, starting with the rth element ($0 \leq r < q$):

$$A_{q;r}(z) = \sum_{n\geq 0} f(nq+r)z^{nq+r}.$$

Set $m = nq + r$. Note that

$$\begin{aligned} A_{q;r}(z) &= \sum_{m\geq 0}[\![q \mid (m-r)]\!]f(m)z^m \\ &= \sum_{m\geq 0}\frac{1}{q}\sum_{k=0}^{q-1}\omega^{(m-r)k}f(m)z^m \\ &= \frac{1}{q}\sum_{k=0}^{q-1}\sum_{m\geq 0}f(m)z^m\omega^{mk}\omega^{-rk} \\ &= \frac{1}{q}\sum_{k=0}^{q-1}\omega^{-rk}A(z\omega^k). \end{aligned}$$

An entirely analogous argument in the multidimensional case gives us the following lemma.

Lemma A.1. *Let $A(z_1, z_2, \ldots, z_m)$ be the ordinary generating function of $f(n_1, n_2, \ldots, n_m)$. Define*

$$A_{q;r_1,\ldots,r_m}(z_1, z_2, \ldots, z_m) = \sum_{n_1 \geq 0} \cdots \sum_{n_m \geq 0} f(n_1q+r_1, \ldots, n_mq+r_m) z_1^{n_1q+r_1} \cdots z_m^{n_mq+r_m},$$

where each $r_i \in \mathbb{Z}_q$. Then

$$A_{q;r_1,\ldots,r_m}(z_1, z_2, \ldots, z_m) = \frac{1}{q^m} \sum_{\nu_1=0}^{q-1} \cdots \sum_{\nu_m=0}^{q-1} \omega^{-(\nu_1 r_1 + \cdots + \nu_m r_m)} A(\omega^{z_1\nu_1}, \ldots, \omega^{z_m\nu_m}).$$

Recall that

$$B(z) = (1+z)^n = \sum_{r=0}^{n} \binom{n}{r} z^r.$$

Substituting $z = 1$ into $B_{q;r}(z)$ we obtain

$$\text{(A.1)} \qquad \sum_{j \equiv r(q)} \binom{n}{r} = \frac{1}{q} \sum_{j=0}^{q-1} \omega^{-rj} (1 + \omega^j)^n.$$

Introduce the notation

$$M_q(n; r_1, r_2, \ldots, r_m) = \sum_{\substack{\nu_0+\nu_1+\cdots+\nu_m=n \\ \nu_1 \equiv r_1(q), \ldots, \nu_t \equiv r_m(q)}} \binom{n}{\nu_0, \nu_1, \ldots, \nu_m}.$$

Plugging $z_1 = z_2 = \cdots = z_m = 1$ into the ordinary generating function $(1 + z_1 + \cdots + z_m)^n$ for the multinomial coefficients, we obtain the following lemma, which generalizes (A.1).

Lemma A.2. *For all $q \geq 2$, $n \geq 0$, and $r_i \in \mathbb{Z}_q$,*

$$M_q(n; r_1, r_2, \ldots, r_m) = \frac{1}{q^m} \sum_{\nu_1=0}^{q-1} \cdots \sum_{\nu_m=0}^{q-1} \omega^{-(\nu_1 r_1 + \cdots + \nu_m r_m)} (1 + \omega^{\nu_1} + \cdots + \omega^{\nu_m})^n.$$

(Note: For the binomial case, see [3, vol. 1, ex. 38, p. 70]. Knuth attributes the roots of unity formula to C. Ramus, 1834.)

Acknowledgments. We wish to thank Nate Kube for help with programming and for helpful discussion. We also thank the referee for carefully reading the paper and suggesting the polynomial based approach that we now adopt.

REFERENCES

[1] P. J. Cameron, *Combinatorics: Topics, Techniques, Algorithms*, Cambridge University Press, Cambridge, UK, 1994.

[2] K. Cattell, F. Ruskey, C. R. Miers, J. Sawada, and M. Serra, *The number of irreducible polynomials over GF(2) with given trace and subtrace*, J. Combin. Math. Combin. Comput., 47 (2003), pp. 31–64.

[3] D. E. Knuth, *The Art of Computer Programming, Volume* 1: *Fundamental Algorithms*, 3rd ed., Addison–Wesley, Reading, MA, 1997.

[4] D. E. Knuth, R. L. Graham, and O. Patashnik, *Concrete Mathematics*, Addison–Wesley, Reading, MA, 1989.

[5] R. Lidl and H. Niederreiter, *Introduction to Finite Fields and Their Applications*, Cambridge University Press, Cambridge, UK, 1994.

[6] C. R. Miers and F. Ruskey, *Counting strings with given elementary symmetric function evaluations* II: *Circular strings*, SIAM J. Discrete Math., to appear.

[7] H. S. Wilf, *What is an answer?*, Amer. Math. Monthly, 89 (1982), pp. 289–292.